Classification Using Transfer Learning on Structured Healthcare Data

by

Akram Farhadi

(Under the Direction of John A.Miller)

ABSTRACT

Recently deep learning has been used as a new classification platform and has been applied to many domains. In some domains such as bioinformatics and healthcare constructing a large-scale well annotated data-set is very difficult. As such labeled data are limited. This structured data in healthcare are small data-set and because of that deep learning approaches do not perform well on their classification. Transfer learning relaxes the hypothesis that learning should occur purely based on specific data-sets, which motivates us to use transfer learning to solve the problem of insufficient training data. In this dissertation, I introduce my efforts toward creating a complete, fully automated and efficient deep transfer learning method to handle the imbalanced data of breast cancer. I compared our results with state-of-the-art techniques for addressing problems of imbalanced learning, poor performance learning and confirmed the superiority of the proposed methods. I conducted a meta-analysis to analyze the status of healthcare-related Transfer Learning(TL) studies in terms of the study targets, TL model(s) used, Healthcare data, type of study area, and level of classification accuracy achieved. Subsequently, a detailed review is conducted to describe/discuss how TL has been applied for improving the accuracy of diagnosis in healthcare including images, text, audio, video and structured Electronic Health Record data classification. I further

present my deep transfer learning model to improve the accuracy of classification in diabetes disease. Finally, I demonstrate the significant performance gains of our model compared to state of art techniques for classification. Based on the experimental results, we concluded that the proposed deep transfer learning on structured data can be used as an efficient method to handle imbalanced classes and poor performance learning on small dataset problems in clinical research.

INDEX WORDS: Deep transfer learning, Class imbalance, Healthcare

Classification Using Transfer Learning on Structured Healthcare Data

by

Akram Farhadi

B.S., Tehran University of Medical Sciences, Iran, 2010

M.S., Tehran University of Medical Sciences, Iran, 2012

A Dissertation submitted to the Graduate Faculty

of The University of Georgia in Partial Fulfillment

of the

Requirements for the Degree

Doctor of Philosophy

Athens, Georgia

2020

Classification Using Transfer Learning on Structured Healthcare Data

by

Akram Farhadi

Major Professor:	John A.Miller
Committee:	Jaewoo Lee
	Ping Ma

Electronic Version Approved:

Ron Walcott
Interim Dean of Graduate School
The University of Georgia
August 2020

For My Family

ACKNOWLEDGEMENTS

I would like to express my most sincere gratitude to my academic and life mentor, Dr. Miller, for all his efforts on me to finally bring me here, completing the Ph.D. study with this dissertation.

I would like to express my special thanks to my Ph.D. committee: Dr. Jaewoo Lee and Dr Ping Ma. Their expertise in deep learning and statistics has always been an invaluable contribution to my achievements in combining deep learning with healthcare data analysis.

Additionally, I would like to Mayo-clinic as breast cancer research of this work is supported by the Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Contents

ACKNOWLEDGEMENTS **v**

List of Figures **viii**

List of Tables **x**

1 INTRODUCTION **1**

1.1 Structured Electronic Health Record 1

1.2 Motivation and research objectives 2

1.3 Contributions and outlines 3

1.4 Dissertation organization 5

2 LITERATURE REVIEW OF DEEP TRANSFER LEARNING IN HEALTHCARE **6**

2.1 Introduction 6

2.2 Deep Transfer Learning in Healthcare 10

2.3 Notations and Definitions of deep transfer learning 11

2.4 A Categorization of Transfer Learning Techniques in Healthcare 15

2.5 Transfer learning in Healthcare through instances 16

2.6 Transfer learning in Healthcare through features 18

2.7 Transfer learning in Healthcare through Neural Network 21

2.8 Transfer learning in Healthcare through adversarial learning 22
2.9 Conclusion 23

3 CLASSIFICATION USING DEEP TRANSFER LEARNING ON STRUCTURED HEALTHCARE DATA 26

Abstract 27
3.1 Introduction 27
3.2 Challenges in Applying Classification in Healthcare 32
3.3 Techniques to Address the Challenges 36
3.4 Related work 45
3.5 Deep Transfer Learning on Structured Data 50
3.6 Baseline Methods for Imbalanced Classification 54
3.7 Training, Validation, and Evaluation 54
3.8 Case Study 1 55
3.9 Discussion 62
3.10 Case Study 2 65
3.11 Discussion 71
3.12 Conclusion 72
3.13 Acknowledgements 73

4 DISCUSSION AND CONCLUSION 79

Appendices 81

A APPENDIX 81

A.1 KL Divergence of original data 82

REFERENCES 86

List of Figures

2.1 Transfer Learning 12

2.2 Deep Transfer Learning for Breast cancer Classification 12

2.3 Transfer Learning 24

3.1 Illustration of transfer learning. Left: locating transferable hidden layers. Right: transferring layers that can improve the learning of the target predictive function. 59

3.2 Deep transfer learning for breast cancer classification 59

3.3 Performance of models on the UCI Mammographic mass data with simulated imbalanced training data distribution 61

3.4 Source encoder 66

3.5 Target encoder 66

3.6 DNN model 66

3.7 Two-sample KL-divergence testing. 78

3.8 BeforeTransfer 78

3.9 AfterTransfer 78

A.1 BeforeTransfer 82

A.2 AfterTransfer 82

A.3 Source encoder 83

A.4 Target encodere 83

A.5 DNN model 83

List of Tables

2.1 Disease classification using Transfer learning 16

3.1 Statistical analysis of target MMHS data (Logistics Regression) 74
3.2 Statistical analysis of target MMD data (Logistics Regression) 75
3.3 Parameters of the Breast Cancer Network 75
3.4 Hyperparameters of the Breast Cancer Network 75
3.5 DTL models 76
3.6 Performance of models on hold out fold of MMHS 76
3.7 Features of Diabetic Readmission Dataset 77
3.8 Parameters of the Diabetes Network using Feature based Transfer learning . 77
3.9 Hyperparameters of the Diabetes network 77
3.10 Performance of models on diabetes dataset 78

A.1 Descriptive statistics of combined WDBC and WPBC source data (Logistics Regression2 84
A.2 Features of WDBC, WPBC and Combined Source Data 85
A.3 Tuned Hyper-Parameters of classifiers on diabetes dataset 85
A.4 Grid search on Neural network hyperparameters 85

Chapter 1

INTRODUCTION

1.1 Structured Electronic Health Record

EHR adoption has facilitated clinical documentation data for research and better inform clinical decision making [1]. However, it was designed primarily for patient care and reimbursement purposes and often lack the standardization necessary for secondary data analysis. Structured data entry can be time-consuming, and its adoption varies widely among different end-users [2],[3]. Clinicians often are unwilling to accept the data entry burden because of the negative impact on physician productivity unless receiving significant returns for their efforts. Therefore insufficiency of labeled data in healthcare renders many statistical approaches unusable. For example, a patient diagnosed with "type 2 diabetes mellitus" can be identified by laboratory values of hemoglobin A1C greater than 7.0, presence of 250.00 ICD-9 code, "type 2 diabetes mellitus" mentioned in the free-text clinical notes, and so on. On the other hand healthcare data suffers from "rare cases" or "rare classes" [4],[5],[6]. Imbalanced property of healthcare data is a major challenge. For example, most health datasets usually have very few cases of the target disease, when compared to the number of healthy patients in the dataset [7]. When positive instances or target class has significantly fewer observations

than negative instances of other classes, the imbalanced classification problem presents in binary classification. The former is usually called a minority class, and the latter, a majority class.

1.2 Motivation and research objectives

Machine learning and deep learning have found many applications in healthcare domains, where we look to build predictive models based on labeled training data.

The success of predictive algorithms largely depends on feature selection and data representation. Contrary to traditional machine learning approaches, deep learning does not require domain-specific data pre-process and it is expected to change human life in the future [8]. Deep learning can help clinicians diagnose disease, identify cancer sites, recognize drug effects and understand the relationship between genotypes and phenotypes, phenotyping and predict infectious disease outbreaks with high accuracy [9],[10]. Phenotyping has many applications in cohort construction for genomic studies, quality improvement, risk adjustment and detection undiagnosed disease [11].

In practice, obtaining high-quality labels are typically unavailable during delivery of care, and to label, new data is time-consuming and expensive, because the domain experts are highly trained physicians. Labels in the electronic health record are scarce, and even when we obtain a training dataset by paying an expensive price, it gets easily out of date and thus can not be effectively applied in the new task [12],[13]. A small and labeled dataset for a specific task is easier to collect but results in machine learning algorithms that tend to perform poorly on new data [14]. On the other hand, leveraging data from other institutes are difficult because of differences with patients and its data coding and capture.

The scarcity of labeled data in healthcare renders many statistical approaches like deep learning unusable [15] because successful training deep learning models always need large

amounts of data to understand the latent patterns of data [16],[17],[18]. In this research, we use transfer learning to mitigate imbalanced classes and small training data problem. The transfer learning methodology relaxes the assumption that the training and test data must have the same feature space or distribution, thus models trained on one domain (source) can be applied to a different domain (target) [19],[20]. We explore the transfer of knowledge from one or more source structured EHR in which training labels are plentiful and more meaningful to target tasks in which have fewer labels with limited clinical value. In particular, we use case studies including breast cancer and diabetes which are two rapidly increasing diseases in the world to illustrate the importance of Deep Transfer Learning on structured EHR data. For case study 1, I used publicly available breast cancer datasets to develop source model and used that to facilitate learning on MayoClinic Health system target DTL classification model. Then I used Mammographic Mass dataset to generalize my results. For case study 2, I used the diabetic readmission dataset to develop a source model. Then, I used this source model to facilitate learning on the Pima Indian diabetes DTL classification model.

1.3 Contributions and outlines

To be specific, two main problems have been addressed in this dissertation. In the following paragraphs, I first explain each problem, then the contributions of this study to resolve those problems are discussed. The first problem is the highly imbalanced data in healthcare which causes classifiers to classify instances as negative and makes classifiers insufficient to get used in clinical decision making. This problem is presented and addresses through in case study 1.

The second problem is the lack of sufficient labeled data for deep learning to accurately predict disease and help clinicians in clinical decision making. Unlike traditional machine

learning methods, in which the creator of the model has to choose and encode features ahead of time, deep learning enables a model to automatically learn features that matter. In this way, a deep learning model learns a representation of the dataset. The pre-order layers in the model can identify high-level features of training data, and subsequent layers can identify the information needed to help make the final decision. Deep learning models extract important features by iteratively transforming the data, "going deeper" toward meaningful patterns in the dataset with each transformation. However, a large dataset is needed for the utilization of complex healthcare data. This problem is presented and addresses through in case study 2.

I proposed a network-based deep transfer learning (DTL) which is non-linear classification model on structured healthcare data to address a class imbalance problem and insufficiency of labeled data in healthcare. Deep learning extracts more effective features by building deep structures. These features relax mismatch between source and target domains. Consequently, deep learning can generate more domain-invariant features for knowledge transfer between domains.

Using transfer learning can potentially solve lack of sufficient labeled data problem by bridging the source and target domains vis-a-vis learning invariant feature representations from the source domain. In transfer learning the source domain can differ from the target domain by having a different feature space, a different distribution of instances in the feature space, or both. This DTL assumes that all the data lie in the same feature space with differences in underlying distributions across tasks. It uses learning from one task on to another task without the requirement of learning from scratch and in this way it directly addresses the smarter parameter initialization point for training the neural network. Using transfer learning, especially in the Healthcare area which suffers from the imbalanced dataset and small labeled dataset improves AUC. AUC has a meaningful interpretation of disease classification from healthy subjects.

1.4 Dissertation organization

Chapter 1 provides an introduction to this study. Chapter 2 provides challenges, current solutions to those challenges, related work in applying transfer learning on structured healthcare data and details of implementing DTL in structured healthcare using two real world case studies. Chapter 3 focuses on the Meta-Analysis of Transfer Learning in Healthcare Effectiveness. In chapter 4, the summary of this study is discussed. It is followed by future directions for this research.

Chapter 2

LITERATURE REVIEW OF DEEP TRANSFER LEARNING IN HEALTHCARE

2.1 Introduction

Electronic health records (EHRs) are capturing the thoughts and orders of the best-trained physicians along with images and reactions from their treated patients. Meanwhile advances in deep learning are beginning to supplement clinical medicine. Machine learning and deep learning have become a new trend in healthcare and opened a research era. Deep learning is fast becoming a key instrument in artificial intelligence applications. For example, in areas such as computer vision, natural language processing, and speech recognition, deep learning has been producing remarkable results. Therefore, there is a growing interest in deep learning. One of the areas where deep learning excels is image classification. With deep learning, researchers discover nonlinear relationships between variables and help clinicians and patients with an objective and personalized definition of disease and solutions. Deep

learning models are basically made up of data itself rather than requiring domain experts and models divide the clinical trials into subgroups according to their clinical information. Deep learning is assisting medical professionals and researchers in discovering the often hidden opportunities of research and decision making within large stores of medical data and, therefore, in serving the healthcare industry better. Deep learning in healthcare provides doctors with accurate disease analysis and helps them treat their patients better, also resulting in better medical decisions. Deep learning algorithms attempt to learn high-level features from mass clinical data, which makes deep learning surpass traditional machine learning. It can automatically extract data features by unsupervised or semi-supervised feature learning algorithms and hierarchical feature extraction. In deep learning models, pre-ordered layers in the model can identify high-level features of training data, and the subsequent layers can Identify the information needed to help the model to make a classification [21].

In contrast, traditional machine learning features need to be designed manually, which increases user burden. Therefore, plenty of experimental works have implemented deep learning models for health informatics, reaching alternative techniques that have been used by most clinicians in order to predict and diagnose disease before actual healthcare problems occur. Nevertheless, the application of deep learning to health information raises a serious problem of the lack of sharable data in healthcare. Deep learning has a very strong dependence on large training datasets compared to traditional machine learning methods because it needs a large amount of data to understand the latent patterns. In deep learning, there is an interesting phenomenon where the scale of the model and the size of the required data has an almost linear relationship. An acceptable explanation for why deep learning requires a large training dataset is that for any given problem, the expressive space of the model must be large enough to discover the patterns under the data.

Although there are no hard guidelines about the minimum number of training sets, including more data can make more stable and accurate models. However, insufficient training

data is an inescapable problem in the medical domain. There is often a lack of available data because in general, the number of patients is limited in a clinical scenarios for most diseases, and particularly for rare diseases, such as certain age-related diseases. In addition, health informatics requires domain experts more than any other domain to label complex data and test whether the model performs well and is practically usable with such small training datasets [22].

The collection of data is complex and expensive that makes it extremely difficult to build a large-scale, high-quality labeled dataset. Although labels generally help to have good performance of clinical outcomes or actual disease phenotypes, label acquisition is expensive. For example, each sample in the bioinformatics dataset is often demonstration a clinical trial or a pain patient. In addition, even we obtain a training dataset by paid an expensive price, it is very easy to get out of date and thus cannot be effectively applied in the new tasks. Therefore, there is still no complete knowledge of the causes and progressions of many diseases due to the lack of available clinical data.

The basis for achieving an accurate model is the availability of large amounts of data with well-structured data store system guidelines. Also, we need to attempt to label EHR data implicitly with unsupervised, semi-supervised, and transfer learning, as previous articles [23]. In general, the first admission patient, disabled or transferred patient may be in worse health status and emergent circumstance, but with no information about medication allergy or any history. If we can use simple tests and calculate patient similarity to see the potential for each risk factor, modifiable complications and crises will be reduced.

Transfer learning relaxes the hypothesis that the training clinical data must be independent and identically distributed (i.i.d.) with the test clinical data, which motivates us to use transfer learning against the problem of insufficient training clinical data [24]. For example, adversarial transfer attempts to handle non-IID data, such as data that is not independent and identically distributed [25]; for instance, social network data utilizes adversarial based

transfer techniques. In transfer learning, the training data and test data are not required to be i.i.d., and the model in the target domain does not need to train from scratch, which can significantly reduce the demand for training clinical data and training time in the target domain. Furthermore, to train the target disease using different disease data, especially when the disease is class imbalanced, transfer learning, multi-task learning, reinforcement learning, and generalized algorithms can be considered. In addition, data generation and reconstruction can be other solutions besides incorporating expert knowledge from medical bibles, online medical encyclopedias, and medical journals.

With the growth of deep learning [23], transfer learning has become integral to many applications especially in medical imaging, where the present standard is to take an existing architecture designed for natural image datasets such as IMAGENET, together with corresponding pre-trained weights (e.g. ResNet, Inception), and then fine-tune the model on the medical imaging data.

Deep learning is well suited for medical data as it can identify patterns in sparse, noisy clinical data, and requires little input-feature engineering. Due to the dominant position of deep learning in Healthcare, a survey on deep transfer learning in the healthcare and its applications is particularly important [26],[27]. This survey paper aims to, define deep transfer learning in healthcare and review transfer learning papers on different types of disease. This paper provides an overview of current methods being used in the field of transfer learning as it pertains to data mining tasks for classification. Then we categorize transfer learning into four main categories and provide a researcher interested in transfer learning in healthcare with an overview of related work. The selected surveyed works in this paper are meant to be diverse and representative of transfer learning solutions in the past 5 years. Most of the surveyed papers provide a generic transfer learning solution; however, some surveyed papers provide solutions that are specific to individual applications.

Since the publication of the transfer learning survey paper by Pan [28] in 2010, there have

been over 700 academic papers written addressing advancements and innovations on the subject of transfer learning in healthcare. These works broadly cover the areas of new algorithm development, improvements to existing transfer learning algorithms, and algorithm deployment in new application domains.

The remainder of this paper is organized as follows. The "Definitions of transfer learning" section provides definitions and notations of transfer learning. The "Homogeneous transfer learning" and "Heterogeneous transfer learning" sections provide solutions to homogeneous and heterogeneous transfer learning, while "Negative transfer" section provides information on negative transfer as it pertains to transfer learning. "Transfer learning application" section provides examples of transfer learning applications. Lastly, "Conclusion and discussion" section summarizes and discusses potential future research work.

2.2 Deep Transfer Learning in Healthcare

Traditional learning in the healthcare area is isolated and occurs purely based on specific tasks, healthcare datasets, and training isolated models. No knowledge can be transferred from one model to another [29],[30]. In transfer learning in healthcare, we can leverage knowledge (features, weights, etc.) from previously trained models on open-source data for training newer models on the specific disease and even tackle problems like having less data for the newer task. These methods as shown in Figure 2.1 benefit the pretrained models in. healthcare. In transfer learning, the neural network is trained in two stages: 1) pretraining, where the network is generally trained on a large-scale benchmark dataset representing a wide diversity of labels/categories; and 2) fine-tuning, where the pre-trained network is further trained on the specific target task of interest, which may have fewer labeled examples than the pretraining dataset [31],[32]. The pretraining step helps the network learn general features that can be reused on the target task. This kind of two-stage paradigm shown in

Figure 2.2, has become extremely popular in many settings, and particularly so in healthcare. This process is generally meaningful and making a significant improvement when the target dataset is small to train and researchers intend to avoid overfitting.

Usually, after training a base network on clinical data, the first n layers are copied and used for the target network and the remaining layers of the target network are randomly initialized. The transferred layers can be left as frozen or fine-tuned, which means either locking the layers so that there is no change during training the target network or backpropagating the errors for both copied and newly initialized layers of the target network [33]. A data augmentation technique [34] is being used to amplify the clinical data in the preprocessing level of transfer learning on images in healthcare. In the context of healthcare images, this involves making a number of non-exact copies, or transformations, of each image. This served to provide the target model with more training examples by incorporating the salient features in multiple orientations.

2.3 Notations and Definitions of deep transfer learning

In this section, we introduce some notations and definitions that are used in this survey. First of all, we give the definitions of a "domain" and a "task", respectively. For a given domain $D = \mathbf{X}, \mathbf{P}(\mathbf{X})$ where $\mathbf{X}$ is a feature random vector and $P(X)$ is its probability, a task T is defined by two components, a label random variable Y, and a predictive function $f(.)$ denoted as $T = Y, f(.)$ which is not observed but learned from a set of feature vector and label pairs x_i, y_i where $\mathbf{x_i} \in \mathbf{X}$ and $y_i \in Y$.

In our cancer classification example, Y is the set of all labels, which is True, False for a binary classification task, and y_i is "True" or "False", and f(x) is the learner that predicts the label value for the patient x. From a probabilistic viewpoint, $f(x)$ can be written as $P(y|x)$. We consider source healthcare domain clinical data as D_S where $D_S = \{(x_{S_1}, y_{S_1}), ..., (x_{S_n}, y_{S_n})\}$,

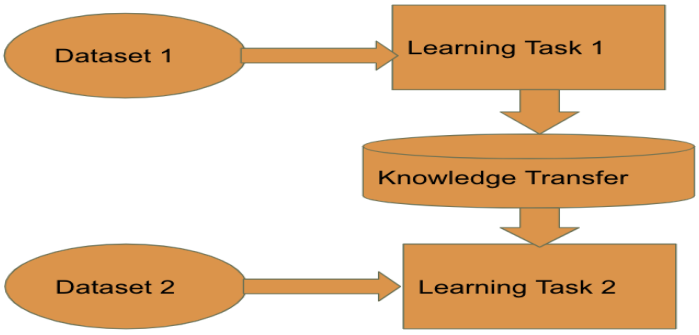

Figure 2.1: Transfer Learning

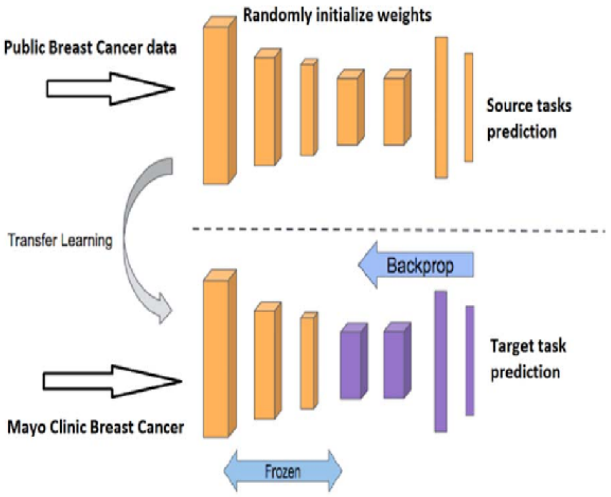

Figure 2.2: Deep Transfer Learning for Breast cancer Classification

where $x_{S_i} \in X_S$ is the ith clinical data instance of D_S and $y_{S_i} \in Y_S$ is the corresponding class label for x_{S_i}. In the same way, D_T is defined as the target domain clinical data where $D_T = \{(x_{T_1}, y_{T_1}), ..., (x_{T_n}, y_{T_n})\}$, where $x_{T_i} \in X_T$ is the ith clinical data instance of D_T and $y_{T_i} \in Y_T$ is the corresponding class label for x_{T_i}. Further, the source task is notated as T_S, the target task as T_T , the source predictive function as $f_{S(.)}$, and the target predictive function as $f_{T(.)}$. Given a source domain D_S and learning task T_S, a target domain D_T and learning task T_T, transfer learning aims to help improve the learning of the target predictive function $f_T(.)$ in D_T using the knowledge in D_S and T_S, where $D_S \neq D_T$, or $T_S \neq T_T$. Condition where $D_S \neq D_T$ means that $X_S \neq X_T$ and/or $P(X_S) \neq P(X_T)$.

The case of traditional machine learning is $D_S = D_T$ and $T_S = T_T$ .On the other hand, the case where $X_S \neq X_T$ with respect to transfer learning is defined as heterogeneous transfer learning. The case where $X_S = X_T$ with respect to transfer learning is defined as homogeneous transfer learning. Going back to the example of cancer, heterogeneous transfer learning [35] is the case where the source cancer has different variables (features) than the target cancer. Alternatively, homogeneous transfer learning [36] is when the cancer metrics are the same for both the source and the target cancer disease. Continuing with the definition of transfer learning, the case where $P(X_S) \neq P(X_T)$ means the marginal distributions in the input spaces are different between the source and the target domains. Referring to the cancer classification, an example of marginal distribution differences is when the source cancer is related to breast and the target cancer is related to skin. Another possible condition of transfer learning (from the definition above) is $T_S \neq T_T$, and it was stated that $T = \{y, F(.)\}$ or to rewrite this, $T = \{Y, P(Y|X)\}$. Therefore, in a transfer learning environment, it is possible that $Y_S \neq Y_T$ and/or $P(Y_S|X_S) \neq P(Y_T|X_T)$. The case where $P(Y_S|X_S) \neq P(Y_T|X_T)$ means the conditional probability distributions between the source and target domains are different. An example of a conditional distribution mismatch is when a particular cancer yields different outputs in the source and target domains. The case of

$Y_S \neq Y_T$ refers to a mismatch in the class space. An example of this case is when the source cancer has a binary label space of true for having breast cancer and false for not having breast cancer, and the target domain has a label space that defines four levels of Birads changes in Breast cancer. Another case that can cause discriminative classifier degradation is when $P(Y_S) \neq P(Y_T)$, which is caused by an unbalanced labeled clinical data set between the source and target domains.

In addition, when there exists some relationship, explicit or implicit, between the feature spaces of the two domains, we say that the source and target domains are related.

Negative transfer, with regards to transfer learning, occurs when the information learned from a source domain has a detrimental effect on a target learner. More formally, given a source domain D_S, a source task T_S, a target domain D_T , a target task T_T , a predictive learner $f_{T1}(.)$ trained only with D_T , and a predictive learner $f_{T2}(.)$ trained with a transfer learning process combining D_T and D_S, negative transfer occurs when the performance of $f_{T1}(.)$ is greater than the performance of $f_{T2}(.)$. The section of negative transfer explains the need to quantify the amount of relatedness between the source domain and the target domain and making decision about the possibility of transferring knowledge from the source domain. Extending the definition above, positive transfer occurs when the performance of $f_{T2}(.)$ is greater than the performance of $f_{T1}(.)$.

Phrases such as transfer learning and domain adaptation are used to refer to similar processes. The following definitions will be used in this paper. Domain adaptation, as it pertains to transfer learning, is the process of adapting one or more source domains for the means of transferring information to improve the performance of a target learner. The domain adaptation process attempts to alter a source domain in an attempt to bring the distribution of the source closer to that of the target.

2.4 A Categorization of Transfer Learning Techniques in Healthcare

There are different strategies and implementations for solving a transfer learning problem. The majority of the homogeneous transfer learning in healthcare solutions apply one of three general approaches which include trying to correct for the marginal distribution difference in the source, trying to correct for the conditional distribution difference in the source or trying to correct both the marginal and conditional distribution differences in the source healthcare domain [37]. The majority of the heterogeneous transfer learning solutions are focused on adjusting the input spaces of the source and target healthcare domains with the assumption that the domain distributions are the same. If the healthcare domain distributions are not equal, then another domain adaptation steps are needed. One more important form of a transfer learning is the scheme of healthcare information transfer referring to what is being transfered. In healthcare area, studies used different types of transfer learning for classifying different types of disease as shown in table 2.1.

The aspect of information transfer is categorized into four general Transfer learning categories including: Transfer learning in Healthcare through instances, Transfer learning in Healthcare through features, Transfer learning in Healthcare through neural Network,Transfer learning in Healthcare through adversarial learning.

Table 2.1: Disease classification using Transfer learning

Disease	$clinical_Data$	Model	Source	Target	Type	Acc/sens(%)	Reference
Cancer							
Skin	Image	DCNN+softmax	AlexNet	ph2 dataset	Network	98.61/98.33	[38]
-	Image	CNN+SVM	MatConNet VGG-M	Interactive Atlas of Dermoscopy	Feature based	73.2	[39]
-	Image	DCNN+SVM	U-Net	PH2	Feature based	93	[40]
Breast	Image	CNN+LR	VGG16,VGG19,ResNet50	BreakHis dataset	feature-based	92.60	[41]
-	Image	CNN+FC	VGG16, ResNet50, InceptionV3	CBIS-DDSM,INbreast	Network	84.16	[42]
-	Image	CNN	AlexNet	CBIS-DDSM	Network	82.6/ 79.10	[43]
-	Image	DCNN	InceptionV3,InceptionResNetV2,Xception,VGGNet	ICIAR 2018 Grand Challenge	Network	92.50	[44]
-	Image	DCNN	ImageNet DCNN	DDSM	Network	82.2	[45]
-	Image	CNN+FC	ALEXnET	dbt DATA	Network	91.3	[46]
-	Image	CNN	GoogleNet,AlexNet	BCDR	network	88	[47]
-	Image	CNN+FC	InceptionV3	BCDR-F03	Network	97.50	[48]
-	Image	CNN+SVM	InceptionV3,VGG19	OASBUD	Feature based	87.82/77.70	[49]
-	Image	CNN	Googles Inception-V3,ResNet50	BACH 2018	network	97.50,91.25/99.9,98	[50]
-	Image	CNN+classifier	InceptionV3	Camelyonn16	Feature based	84	[51]
Prostate	Image	DCNN+FC	InceptionV3,VGG16	prostateX from 3TMR,MAGNETOM Trio		93.4	[52]
-	image	CNN+FC	Alex-Net, GoogleNet,VGGNet	PROSTATEx-2	Network	86.92/88.09	[53]
Lung	Image	CNN+FC	GoogleNet	Network	LIIDC data	81/84	[54]
Cervix	Image	CNN	VGG16 on imageNet	42 cervical cancer patients	Network	89/75	[55]
Colon	Images	DCNN+FC	Inception-V3,VGG16,SE-Resnext50	CLM images	Network	91.2/82.8	[56]
-	Image	CNN+SVM	VGG16	CC-i-Scan Datatbse	Feature based	87.70	[57]
-	Image	CNN+SVM	AlexNet,VGG16,VGG19,GoogleNet	MICCAI 2015	Feature based	92	[58]
-	Image	CNN	Inception-V3	Harbin medical hospital	network	94.4/85	[59]
-	Image	CNN	VGG-16	EBRT,BT	Network	70/61	[55]
Gastric	Image	DCNN+FC	VGG16, InceptionV3,InceptionResNetV2	M-NBI images	Network	98/98	[60] -
Alzhimer	Image	DCNN+last-3L-finetuned	AlexNet	(OASIS) dataset	Network	92.85/92.85	[61]
-	Image	CNN+FC	VGG-16	ADNI dataset	Network	95.73	[62]
-	Image	TrAdaboost	ADNI dataset	local hospital	Instance	93.7/87	[63]
-	Image	CAE+3DCNN	CAE	ADNI database	Feature	86.60/88.55	[64]
-	Image	CNN+FC	VGG,Inception	OASIS data	network	92.3,96.25	[65]
-	Image	CNN+FC	AlexNet	OASIS data	network	92.85/92.85	[61]
-	Image	CNN+FC	ADNI dataset trained on similar VGG19	ADNI dataset,	network	99.2	[66]
-	Image	DTFS+SVM	ADNI dataset	ADNI dataset	Feature based	79.4	[67]
-	Image	Multi-Domain Transfer Feature+SVM	(http://adni.loni.usc.edu/)	186 AD, 395 MCI, and 226 NC)	Feature based	94.7	[68]
Diabetes	Image	AlexNet+SVM	AlexNet	SERI dataset	feature based	96.7/97.66	[69]
-	Image	CNN+FC	AlexNet	Kaggle partition (MildDR)	Network	74	[70]
-	Image	CNN+FC	GoogleNet	Duke OCT data	Network	94	[71]
-	Image	CNN	AlexNet,VggNet,GoogleNet	DR1, MESSIDOR datasets	Network	92/86	[72]
-	Image	CNN+FC	AlexNet,GoogleNet,ResNet,VggNet	OCT image database	network	97	[73]
-	Image	DCNN	VGGNet, ResNet	SiDRP	Network	98.9	[74]
Brain	Image	CNN+FC	InceptionV3	ISPECT scans	95/96.3	network	[75]
-	Image	CNN+SVM	AlexNet	PaHaW	Feature based	98.28	[76]
-	Image	DCNN+softmax	VGG CNN	St.Jude childrens hospital	Network	89.8	[77]
-	Image	CNN+SVM	AlexNet,GoogleNet,VGGNet	Figshare	Feature based	98.69	[78]
-	Image	CNN+FC	ResNet34	Brain MR images	network	97.86	[79]
-	Image	FCN	FLAIR and T1	RUN DMC	Network	63	[80]
-	Image	CNN+SVM	GoogleNet	Brain MRI images	Feature based	93.3	[81]
-	Image	CNN+SVM	AlexNet	Aminu Kano Teaching Hospital data	Feature based	93	[82]
Lung	image	DCNN+Fc	ChestX-ray14,VGG16,DenseNet,Xception,InceptionV3	Labeled(oct) and chest X-ray images	Network	96.7/92	[83]
-	Image	DCNN+Softmax	Inception-v3	pulmonary JSRT database	Network	94.71/86.40	[84]
-	structured	PBD	chronic bronchitis,emphysema data	COPD demographic,EMR,Lab data	instance+ feature learning	90.8	[85]
-	Image	DCNN	GoogleNet,AlexNet	ILD dataset	Network	86.90	[86]
-	Image	CNN+FC	VGG16 on Cifar-10	CT images taken at Yamaguchi University Hospital	network	83.8	[87]
-	Image	MIL	Danish Lung cancer screening	DLCST	instance based	90	[88]
-	Image	CNN	ALOT,DTD,KTB,KTH-TIPS-2b	The HUG database	Network		[89]
Heart	Image	CNN+FC	VGG-Net	PTB database	network	99.2/98.76	[90]
-	image	CNN+SVM	InceptionV3,GoogleNet	ICBEB	Feature based	85.8	[91]
-	Image	DCNN	Inception-ResnetV2	IVOCT	Network	78.9/77.9	[92]
-	Image	CNN+FC	InceptionV3,ResNet50,Xception	2058 masses	Network	79	[93]
Glucoma	Image	CNN+FC	VGG16	RIM-ONE	Network	92.4/91.7	[94]
-	Image	CNN,RF,SVM	RS 3000 OCT	Topcon OCT-1000	Feature based	82.5	[95]
-	Image	CNN+FC	InceptionV3	EyeyPacs kaggle	Network	64	[96]
-	Image	CNN	VGG16,InceptionV3,ResNet	OCT images	Network	91/92	[97]
-	Image	CNN	InceptionNetwork	OCT data	Network	98.6/95.6	[98]
Others	activity	CNN_FC	UCI smartphone	UCI smartphone	Network	98	[99]
-	Image	DCNN+FC	VGGNet	Shanghai Jiaotong University	Network	96.6	[100]
-	activity	CNN+MMD	UCI,USC-HAD	UCI,USC-HAD	feature based	87	[101]
-	Image	CNN+SVM	Inception-ResNet-v2	Zenodo repository	Feature based	96	[102]
-	Text	RNN	(http://ww.Sentiment140.com)	(http://www.taoconnect.org)	feature based+Network	78	[103]
-	protein	MIMTL	(http://lamda.nju.edu.cn/files/MIMLprotein.zip)	(http://lamda.nju.edu.cn/files/MIMLprotein.zip)	Instance based	high-rank	[104]
-	Text	BiDirectional LSTM	MIMIC III dataset	ShARe/CLEF disorders	Feature based	86	[105]
-	Image	CNN+FC	VGG16,ResNet50,IncpetionV3	Jeol 1400 TEM	Network	95/95	[106]
-	Image	CNN+FC	InceptionV3	endoscopy image dataset	Network	98/87	[107]

2.5 Transfer learning in Healthcare through instances

Reusing knowledge from the source domain to the target task is usually an ideal scenario. In most cases, the source domain clinical data cannot be reused directly. Rather, there are certain instances from the source domain that can be reused along with target clinical data to improve results. Instance-based transfer learning [108] assumes that there are some parts

of source clinical data that are usable with a few labeled target instances to train a learner for the target task and improve the learning performance [109], [110]. Selecting the instances from the source clinical data that will benefit the target task is a key step for instance-based transfer learning [111]. A common method used in this case is assigning weights to source domain instances so that they can match the target domain well. One of the algorithms used extensively in healthcare for instance based transfer learning is TrAdaBoost[112], [113]. TrAdaBoost extends AdaBoost for transfer learning [114].

AdaBoost[115] is a learning framework which aims to boost the accuracy of a weak learner by carefully adjusting the weights of training instances and learn a classifier accordingly. Paper[63] used TrAdaboost to classify Alzheimer's Disease(AD), mild cognitive impairment(MCI), and normal controls (NC) data from Alzheimer's Disease Neuroimaging Initiative (ADNI). These reweighted instances are then directly used in the target domain for training. These reweighting algorithms work best when the conditional distribution is the same in both domains. Due to the probability distribution discrepancy across healthcare domains, it is natural to account for the difference by directly inferring the resampling weights of instances based on feature distribution matching across the source and target clinical data [116],[117].

Another way to use instance-based transfer learning is to use a combination of source domain classifiers to label the unlabeled target clinical data[118]. This is accomplished by first building a classifier for each separate source domain. Then a weight value is found for each classifier as a function of the closeness in conditional distribution between each source and the target domain[119]. The weighted source classifiers are summed together to create a learning task that will find the pseudo labels (estimated labels later used for training) for the unlabeled target clinical data. Finally, the target learner is built from the labeled and pseudo labeled target clinical data.

2.6 Transfer learning in Healthcare through features

This approach aims to minimize domain divergence and reduce error rates by identifying good feature representations that can be utilized from the source to target domains. Depending upon the availability of labeled data, supervised or unsupervised methods may be applied for feature-representation-based transfers. In a transfer learning environment, there are scenarios where a feature in the source domain may have a different meaning in the target domain. The issue is referred to as context feature bias, which causes the conditional distributions between the source and target domain to be different. Adding duplicate copies of the original feature set in the augmented source feature space represents a common feature set, a source-specific feature set, and a target-specific feature set and solves the problem of context feature bias.

Feature-based transfer learning builds the most intuitive way by finding a good common feature representation known as latent space to minimize the distributions differences while preserving the discriminative ability in both source and target domains [120],[121] which is known as feature-based domain adaptation.

In the context of images, Features extracted by deep neural networks can be seen as complex hierarchical representations of the inputs, which can well capture recessive characteristics inside the images. Once the common feature space is built [99], it does not need to re-train if more source data is available[122],[123]. Depending on the availability of labeled instances in the target domain, feature-based domain adaptation approaches can be divided into two categories: semi-supervised approaches with labeled instances[124],[125],[126] and unsupervised approaches without labeled instances(Self-Taught learning).

Semi-supervised Transfer Learning Semi-supervised learning exploits connections between the input distribution $P(X)$ and a target conditional distribution $P(Y|X)$. In general, these two distributions, seen as functions of x, may be unrelated to each other. But in the world

around us, it is often the case that some of the factors that shape the distribution of input variables X are also predictive of the output variables Y [127]. Deep Learning relies heavily on unsupervised or semi-supervised learning and assumes that representations of X that are useful to capture P(X) are also in part useful to capture $P(Y|X)$.

In [125] authors applied semi-supervised transfer learning to classify Seizure. They use the unlabeled testing clinical data to remedy the shortage of training clinical data. Study[128] use transductive transfer learning where they have a set of labeled training data in the source domain and a set of unlabeled data in the target domain. By minimizing the Maximum mean discrepancy (MMD) [129],[130], the difference in clinical data distribution between the source and the target domains reduced effectively, which made the testing performance close to the training performance. Also, other types of semi-supervised transfer learning have been used to classify disease. Paper [131] used samples of AD and NC as labeled clinical data and MCI-C and MCI-NC as unlabeled clinical data. Then, the SVM model is trained based on both labeled and unlabeled clinical data. Finally, samples of MCI-C and MCI-NC are treated as testing clinical data to evaluate the performance of each learned model.

We note that even though semi-supervised learning was originally defined with the assumption that the unlabeled and labeled data follow the same class labels [132], it is sometimes conceived as "learning with labeled and unlabeled data." Under this broader definition of semi-supervised learning, self-taught learning would be a particularly widely applicable instance of it.

Unsupervised Transfer Learning The source and target domains are similar, but the tasks are different. In this scenario, labeled clinical data is unavailable in either of the domains. In the unsupervised representation-learning phase, one may have access to examples of only some of the classes, and the representation learned should be useful for other classes. One, therefore, assumes that some of the factors that explain $P(X|Y)$ for Y in the training classes,

and that will be captured by the learned representation, will be useful to predict different classes, from the test set. In particular, transfer learning from unlabeled clinical data for predictive tasks is known as self-taught learning [133], where a joint generative model is not assumed to underlie unlabeled samples even though the unlabeled samples should be indicative of a structure that would subsequently help predict tasks.

self-taught learning places fundamentally fewer limitation on the type of unlabeled clinical data[134], in many possible applications (such as image, audio or text classification) it is much easier to apply than typical semisupervised learning or transfer learning methods. And while we treat any biological motivation for algorithms with great caution, the self-taught learning problem perhaps also more accurately reflects how humans may learn than previous formalisms [135], since much of human learning is believed to be from unlabeled data. self-taught learning consists of two stages: First they learn a representation using only unlabeled data. Then, we apply this representation to the labeled data, and use it for the classification task. Once the representation has been learned in the first stage, it can then be applied repeatedly to different classification tasks; for example, once a representation has been learned from Internet images, it can be applied not only to images of animals but also to medical image classification tasks.

It seems that any algorithm for the self-taught learning problem must, at some abstract level, detect structure using the unlabeled clinical data. Many unsupervised learning algorithms have been devised to model different aspects of "higher-level" structure; however, their application to self-taught learning is more challenging than might be apparent at first blush [136]. Principal component analysis (PCA) is among the most commonly used unsupervised learning algorithms. It identifies a low-dimensional subspace of maximal variation within unlabeled clinical data. Principal component analysis (PCA) is used for optimization and dimensionality reduction. To address the difference in marginal distribution between the domains, the maximum mean discrepancy distance measure could be used to compute the

marginal distribution differences and then get integrated into the PCA optimization algorithm. At the end, the features identified by the modified PCA algorithm could be used to train the final target classifier[137].

2.7 Transfer learning in Healthcare through Neural Network

This approach works on the assumption that the models for related tasks share some parameters or prior distribution of hyperparameters. Unlike multitask learning, where both the source and target tasks are learned simultaneously, for transfer learning, we may apply additional weightage to the loss of the target domain to improve overall performance.

In network based transfer approach, some kind of network based model is supposed and the transferred knowledge is encoded into parameters [103]. Network-based deep transfer learning refers to the reuse the partial network that pre-trained in the source domain, including its network structure and connection parameters, transfer it to be a part of deep neural network which used in target domain[138],[45]. First, network gets trained in source domain with large-scale training dataset. Second, partial of network pretrained for source domain are transfer to be a part of new network designed for target domain[75]. Finally, the transfered sub-network may be updated in fine-tune strategy [55],[100],[139].

Usage of transfer learning and CNN in visual recognition tasks gives better results. If a comparison is made between deep approaches, transfer learning with a pre-trained model method is more successful than end-to-end CNN approaches. The reason for this is the need for large clinical data to obtain an accurate CNN model. Most of the related studies are tested with one or two datasets [65].

Learning to transmit is often faster than training a new neural network because all the parameters in the new network are not estimated from scratch. In the lower layers of the

network, more general features exist such as color and shape and they can be transferred to other tasks as well. However, in higher layers, features are more task-specific [140].
Recent studies in healthcare have used pre-trained CNNs to produce human-level diagnostic results in the classification of cancer, Alzheimer [62],lung, diabetes and heart disease, and other diseases [75]. They demonstrated that transfer learning from deep neural network pre-trained on non-medical data(mostly images) can readily be applied to the analysis of plain medical images. Furthermore, good accuracy of classification can be achieved even with modest sample sizes. It had also been used in text classification. Universal Sentence Embeddings are a huge step forward in enabling transfer learning for diverse NLP tasks. ELMo gives us word embeddings which are learned from a deep bidirectional language model (biLM) [141], which is typically pre-trained on a large text corpus, enabling transfer learning for these embeddings to be used across different NLP tasks.

2.8 Transfer learning in Healthcare through adversarial learning

Unlike the preceding three approaches, the adversarial transfer attempts to handle non-IID data, such as data that is not independent and identically distributed. In other words, data, where each data point has a relationship with other data points; for instance, social network data utilizes adversarial transfer techniques.
Adversarial learning has been successfully embedded into deep networks to learn transferable features, which reduce distribution discrepancy between the source and target domains [142]. Existing domain adversarial networks assume fully shared label space across domains. In the presence of big clinical data, there is a strong motivation for transferring both classification and representation models from existing large-scale domains to unknown small-scale domains.

Adversarial based deep transfer learning methods learn new representations through adversarial learning processes. To learn domain-uninformative representations, [143] used Domain Adversarial Training of Neural Network (DANN) in order to add one domain classifier at the last block and learned domain invariant features by minimizing the loss of this classifier and utilizing reversed gradient during the backpropagation process.

Adversarial-based deep transfer learning refers to introduce adversarial technology inspired by generative adversarial nets (GAN)[144] to find transferable representations that apply to both the source domain and the target domain. It is based on the assumption that "For effective transfer, a good representation should be discriminative for the main learning task and indiscriminate between the source domain and target domain" [145],[146],[147]. Clustering-based approaches achieve transfer learning by building a similarity graph between all instances and the weight on each edge represents the similarity between two instances [148]. On the other hand, fooling the adversarial network to match the distribution of outlier source clinical data and target clinical data will make the classifier more likely to classify target data in these outlier classes, which is prone to negative transfer.

Among adversarial learning-based approaches, most works [149] are based on Generative Adversarial Networks [150] by using generators to synthesize clinical images or representations in different domains to learn domain invariant features. For example development of an algorithm to learn a shared representation for classifying labeled source clinical data and reconstructing unlabelled target clinical data.

2.9 Conclusion

Electronic health records (EHRs) can improve the ability to diagnose diseases and reduce—even prevent—medical errors, improving patient outcomes. The application of deep learning to clinical data from Electronic Health Record is limited by the scarcity of mean-

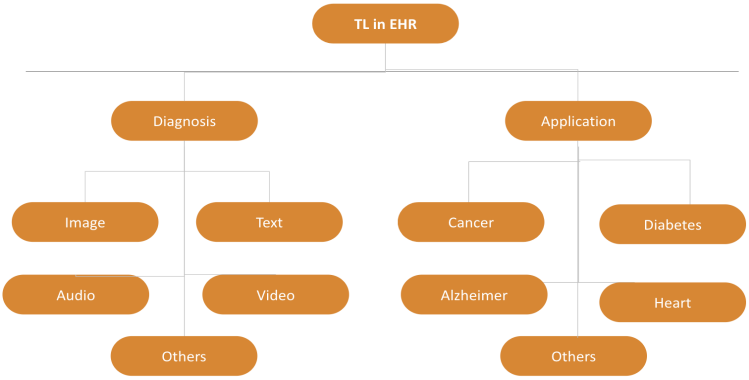

Figure 2.3: Transfer Learning

ingful labels. In this survey paper, we reviewed and categorized current researches of deep transfer learning in the healthcare area. Application of transfer learning in different types of diseases is reviewed and Deep transfer learning in the healthcare area is classified into four categories including instances-based deep transfer learning, feature-based deep transfer learning, network-based deep transfer learning, and adversarial-based deep transfer learning. Transfer learning in healthcare is used in multiple context to make a diagnosis of different types of disease as shown in figure 2.3.

Most current researches focus on supervised learning, how to transfer knowledge in unsupervised or semi-supervised learning by the deep neural network may attract more and more attention in the future. Negative transfer and transferability measures are important issues in traditional transfer learning. The impact of these two issues in deep transfer learning also requires us to conduct further research. Network-based Transfer learning in healthcare is mostly applied in the healthcare area and second-grade feature-based transfer learning is more popular. Instance-based transfer learning has been used less in healthcare and that might be because of the difficulty of adapting instances from other healthcare institutions to improve the accuracy of prediction in the specific healthcare area. Also, a very attractive

research area is to find stronger physical support for transfer knowledge in the deep neural network, which requires the cooperation of physicists, neuroscientists, and computer scientists.

In many transfer learning solutions, the domain adaptation process performed is focused either on correcting the marginal distribution differences or the conditional distribution differences between the source and target domains. Correcting the conditional distribution differences is a challenging problem due to the lack of labeled target clinical data. To address the lack of labeled target clinical data, some solutions estimate the pseudo labels for the target clinical data, which are then used to correct the conditional distribution differences. This method is problematic because the conditional distribution corrections are being made with the aid of pseudo labels. Improved methods for correcting the conditional distribution differences is a potential area of future research.

Additionally, the optimal transfer is another fertile area for future research. Negative transfer is defined as a source domain having a negative impact on a target learner. The concept of optimal transfer is when selected information from a source domain is transferred to achieve the highest possible performance in a target learner. There is an overlap between the concepts of negative transfer and optimal transfer; however, optimal transfer attempts to find the best performing target learner, which goes well beyond the negative transfer concept.

Studies have demonstrated comparable results to the state-of-the-art for automated disease detection having trained the model using only a modest sample size. Deep transfer learning has the potential to significantly improve workflow productivity, minimize the risk of error, and prevent patient harm by reducing diagnostic delays.

Chapter 3

CLASSIFICATION USING DEEP TRANSFER LEARNING ON STRUCTURED HEALTHCARE DATA[1]

[1]Akram Farhadi,David Chen,Rozalina McCoy,Christopher Scott,Celine M. Vachon,Jingyi Zhang,Ping Ma,Che Ngufor and John A. Miller."Classification Using Deep Transfer learning on Structured Healthcare Data" To be submitted to Journal of Data Science and Analytics (JDSA) (2020).

ABSTRACT

The primary limitation of building a supervised learning system for healthcare data is access to a sufficiently large, labeled dataset. A small and labeled dataset for a specific task is easier to collect but results in machine learning algorithms that tend to perform poorly on new data. To address this problem, we propose an accurate and efficient deep transfer learning method to handle the problem with small datasets in healthcare, in particular, an imbalanced data problem. In contrast to existing approaches based primarily on large image databases, we focus on structured data, which has not been commonly used for deep transfer learning. We use several publicly available breast cancer datasets to generate a source model and transfer learned concepts to predict high-grade malignant tumors in patients diagnosed with breast cancer at Mayo Clinic. We then use the diabetes dataset to generalize the idea of transfer learning on structured data. We compare our results with state-of-the-art techniques for addressing problems of imbalanced learning, poor performance learning and demonstrate the superiority of the proposed methods. To further demonstrate the ability of the proposed method to handle different degrees of class imbalance, a series of experiments are performed on publicly available breast cancer data under simulated class imbalanced settings. Based on the experimental results, we conclude that the proposed deep transfer learning on structured data can be used as an efficient method to handle imbalanced class and poor performance learning on small dataset problems in clinical research.

3.1 Introduction

The imbalanced data problem presents a challenge to the machine learning algorithms. These algorithms typically generalize patterns observed over the majority class and ignore those observed over the minority class [151]. The Synthetic Minority Over-Sampling Technique (SMOTE) [152] is currently the most popular approach in handling the class imbalance prob-

lem. SMOTE generates artificial instances of minority classes within the overlapping regions to render the data more balanced. SMOTE has been widely used to solve imbalanced data problems in many medical areas, such as breast cancer classification, prostate cancer staging, and medical imaging [153],[154],[155],[156], [157],[158].

However, SMOTE and variants such as RUSBoost [159] are pre-processing techniques, which may cause the information deficiency problem. SMOTE may generate uninformative data that is useless for training, while RUSBoost may lose informative samples. Recently, SMOTE has been shown to perform better in handling imbalanced data when combined with ensemble learning techniques [156],[160],[161],[162]. However, as demonstrated in this study, these hybrid techniques can still generate sub-optimal classification results for severely imbalanced breast cancer datasets.

Another major problem in most machine learning algorithms, especially deep learning methods, is data dependence. The algorithms require large amounts of training data to obtain well predictive models for the test data. However, in healthcare, it is very difficult to construct a large-scale well-annotated dataset on a particular disease due to the complexity or rarity of the disease, heterogeneity of clinical data sources, and costs associated with annotating the data. Insufficient training data is an inescapable problem in healthcare. Even if sufficient training data can be obtained and annotated, it becomes outdated rapidly due to the influx of new data, which shifts the distributions of the training and testing sets, and makes the model inapplicable.

The transfer learning methodology relaxes the assumption that the training and test data must share the same feature space or distribution, thus models trained on one domain (source) can be applied to a different domain (target) [19],[20]. As a result, transfer learning can be used to mitigate insufficient training data and class imbalance problems in a given target domain. Deep transfer learning (transfer learning with deep learning) has been well

studied in the context of the image classification, i.e. to transfer knowledge from complex publicly available image databases to solve small healthcare image classification problems [20],[163]. However, recent research in [164] found that almost one third of the medical imaging procedures performed in the US and other high-income countries are unnecessary. Using these irrelevant images in training classification models may lead to more imbalanced problem or negative transfer of information in transfer learning settings [20]. Furthermore, imaging procedures are expensive, and may have radiation risks [164]. Among the electronic health records (EHR) [165], images comprise only a small fraction. A large proportion of the EHR are structured data [166]. Structured data, such as diagnosis, medication and laboratory test results are organized in a specific manner. It is used by clinicians and researchers to diagnose and manage diseases. Therefore, transfer learning on large and readily available structured healthcare data can be a cost-effective and computationally efficient learning strategy.

Deep learning extracts more effective features by building up deep structures. These features relax the mismatch between source and target domains. Consequently, deep learning can generate more domain-invariant features for knowledge-transfer between domains. At present, there are many works on combining transfer learning and deep learning. [167] proposed a shared hidden layer multilingual DNN (SHL-MDNN), in which the hidden layer is common in many languages, while the softmax layer is language-dependent.

Deep Transfer Learning (DTL) is largely applied to image data due to numerous public available image datasets for large scale deep learning (e.g. ImageNet [168]) and pre-trained models generated on these image databases (e.g. AlexNet [169]). Although DTL is popular in image analysis, its application on structured data to address the small dataset problem and the imbalanced learning problem is understudied. In this paper, we propose a DTL approach on structured data to address the imbalance and insufficient data classification

problem of healthcare data.

Our central hypothesis is that DTL, also known as domain adaptation, enables structured-data-based solutions that can improve the early detection and classification of disease.

Using two case studies, we illustrate the importance of DTL and its applicability to structured healthcare data. Case study 1, represents the efficiency of transfer learning on a structured breast cancer dataset which is imbalanced. Further using the Mammographic Mass dataset with different imbalanced ratio we generalize our results. Case study 2, represents the importance of transfer learning on improving the accuracy of classification of structured diabetes dataset.

Breast cancer is the most frequently diagnosed cancer among women in the U.S. and worldwide and is a leading cause of cancer-related death in women [170]. Early diagnosis is essential for successful treatment and survival [171]. Thus, the accurate and efficient early diagnosis of breast cancer is critical. Digital mammography is currently the most commonly used and widely available method for detecting masses or abnormalities suggestive of breast cancer [172]. The widespread use of mammography screening has significantly improved breast cancer survival [170],[173]. However, the classification of masses into benign or malignant by most classification methods has been hindered by the rarity of malignant events.

Diabetes is one of the common and rapidly increasing diseases in the world. According to the International Diabetes Federation, there are 285 million diabetic people worldwide. This total is expected to rise to 380 million within 20 years [174]. Diabetes contributes to heart disease, increases the risks of developing kidney disease, nerve damage, blood vessel damage and blindness. So, efficiently predicting diabetes is a critical concern.

To demonstrate the general applicability of the proposed DTL approach, extensive evaluation experiments were performed on a Mayo Clinic breast cancer dataset, the publicly available UCI Mammographic Mass breast cancer dataset and publicly available UCI Pima

Indians diabetes dataset [175]. The evaluations on the UCI Mammographic mass dataset were performed by constructing different class imbalance ratios (IR), defined as the ratio of the number of instances in the majority class to the number of examples in the minority class. Thus, high IRs (¿10) represent severely imbalanced datasets. These simulated experiments allow us to quantify the model performance gains under controlled varying degrees of IRs in the training data. Experimental results show a clear advantage of DTL over 12 state-of-the-arts methods, including standard deep learning methods, XGBoost, SMOTEBoost, and RUSBoost.

All data have hidden features that hold immense predictive power if we can extract them before applying classical machine learning algorithms. Deep learning is a branch of machine learning that extracts features by utilizing multi-layer artificial neural networks with many hidden layers stacked one after the other.

3.1.1 Deep Learning

Recently deep learning has been applied to process EHRs for a specific, usually supervised, predictive clinical task [176]. Deep learning can help clinicians diagnose disease, recognize cancer sites, understand the relationship between genotypes and phenotypes, explore new phenotypes, and predict disease with high accuracy [177],[178],[179]. It discovers nonlinear relationships between features and helps clinicians in clinical decision-making processes. Deep learning is a type of machine learning algorithm that has deeper (or more) hidden layers of similar function cascaded into the network and can learn the representation of healthcare data with multiple levels of abstraction. In contrast to traditional models, its approach does not require domain-specific data pre-processing, and it is expected that it will ultimately improve human life in the future [180].

The deep learning algorithms use simple features in the lower layers and more complex fea-

tures in the higher layers. A neural network with one hidden layer is shallow while the neural network with two or more hidden layers is deep [181]. The hidden layers can be considered as increasingly complex feature transformations and the final output layer as a non-linear classifier making use of the most abstract features computed in the hidden layers. While the non-linear classifier should be different for different datasets, the feature transformations can be shared across related domain datasets.

3.2 Challenges in Applying Classification in Healthcare

As stated in [180],"Nevertheless, the application of deep learning to health informatics raises the number of challenges that need to be resolved, including data informativeness (high dimensionality, heterogeneity, multi-modality), lack of data (missing values, class imbalance, expensive labeling), data credibility and integrity, model interpretability and reliability (tracking and convergence issues as well as overfitting), feasibility, security, and scalability" [180].

In this section we explain key challenges in applying machine learning in healthcare area and how to handle them. In short, health represents a distinct challenge for machine learning because of human still-limited understanding of disease, the effects of human interventions, and the lack of integrated data that can effectively capture this information at meaningful scale. Given this more challenging analytical environment, we need to be more thoughtful about how we employ machine learning in health and healthcare.

3.2.1 Data Scarcity Problem in Healthcare

Healthcare data have several distinguishing characteristics that make them different from data in other areas. Healthcare data are difficult to access due to patient privacy considerations and most researchers in the healthcare area have difficulties practicing data science due

to the risk of data misuse by other parties and lack of data-sharing incentives. Healthcare data are often collected using fixed forms and are relatively structured, partially due to the extraction process that simplifies raw data [182]. Another important feature is that medicine is practiced in safety-critical conditions in which decision-making activities should be supported by explanations. Healthcare data can be costly due to human involvement, the use of expensive instrumentation, and the discomfort of the patients involvement. Healthcare data are relatively small compared to data from other areas and may be collected from a non-reproducible situation. Healthcare data can be further affected by several sources of uncertainty, such as measurement errors, missing data, or errors in coding the information buried in textual reports. Therefore, domain knowledge is more important in both analyzing the data and interpreting the results [183].

Insufficiency of labeled data in healthcare is one of the main problems of machine learning and data mining, because the insufficient size of data is very often responsible for poor performances of learning, how to extract the significant information for inferences is a critical issue [184]. Semi-supervised learning could balance performance and precision using small sets of labeled or annotated data and a much larger unlabeled data collection. The scarcity of labeled data in healthcare renders many statistical approaches like deep learning unusable [15]. There are two possible ways to overcome the data scarcity problem. One is to collect more data while the other is to design techniques that can deal with smaller datasets.

3.2.2 Class Imbalance Learning

An imbalanced dataset introduces a unique challenge to the learning problem. Imbalanced data typically refer to a classification where the number of observations for one class is vastly more than those in the other class. Class imbalance learning is a challenging task mainly because standard classification algorithms assume that the test data are drawn from the

same distribution as the training data. Thus, if the training and testing data samples were drawn from a different distribution, the generated model may produce inferior results. Another reason is that most standard classifiers are designed to optimize a loss function based on minimizing error (or maximizing the predictive accuracy) on the training data. However, the predictive accuracy is a misleading performance metric in imbalanced learning as it assigns an equal cost to false positives and false negatives. When the dataset is imbalanced, for example, 90% of instances are healthy and only 10% have the disease - there is a great way to lower the cost. Predict that most instances belong to the healthy class and get an accuracy of 90%. But, in the case of a rare and fatal disease, the actual costs that we assign to every error are not equal [185]. As a result, failing to diagnose the disease of a sick person is much higher than the cost of sending a healthy person to more tests [186].

Sensitivity is another performance metric that assesses the effectiveness of the classifier on the positive/minority class whereas specificity assesses the classifiers effectiveness on the negative/majority class. Sensitivity is the proportion of true positives that are correctly identified by a diagnostic test. Specificity is the proportion of the true negatives correctly identified by a diagnostic test. Using the Receiver Operating Characteristic (ROC) curve helps us in determining an appropriate balance between sensitivity and specificity. Precision and F-measure are other preferable performance metrics that can better evaluate the minority class. Precision denotes the proportion of predicted positive cases that are real positives. F-measure conveys the balance between precision and sensitivity [187].

Existing methods for dealing with the class imbalance problem can be categorized into three major groups: (a) Data sampling methods: This is a pre-processing step in which the training data is modified to produce a more balanced class distribution, allowing the machine learning algorithm to be capable of learning the classes as in standard classification tasks [151],[187],[188],[189]. This is the most popular imbalanced learning method, and typical

approaches include down-sampling majority classes, oversampling minority classes, or both [152],[153],[155], [156],[159].

In over-sampling techniques, some minority examples are duplicated, while SMOTE methods generate artificial minority examples. Both approaches can lead to the creation of uninformative examples or examples that are correlated (i.e. non-i.i.d), thus making it difficult to learn by the machine learning method [155],[190].

In contrast, down-sampling balances the data by removing majority examples and therefore may lead to information loss. (b) Algorithmic modification: Modifying existing algorithms to be more attuned to the class imbalance problem, e.g. placing more emphasis on the minority classes, improving the information content of the training data, or modifying the decision boundary between classes. (c) Cost-sensitive learning: This approach assigns different costs or weights to the classes, e.g. a higher penalty for misclassifying the minority class samples. The main problem with this approach is that it is often difficult to optimize the cost matrix for a given problem [187].

The majority of studies on imbalanced learning have shown that the significant loss in performance is mainly due to the skewed class distribution, which is given by the IR [190]. However, some studies have also suggested several factors other than class imbalance that may be responsible for performance degradation. For example, feature selection was found to be beneficial in handling high-dimensional imbalanced datasets. The conclusion from [191] was that before feature selection, samples in the original data were too disparate to make generalizations about class memberships. However, when the right features are selected, a significant performance jump was observed, regardless of the classifier used. [191] indicates that the original dataset may lack density and information or that the classes may be overlapped, but feature selection provides the right information needed to discriminate between the classes.

3.2.3 Combining Data Sources

Insufficiency of data in healthcare makes urgent the need to use other datasets in the same area. By combining these datasets with other data sources, it may be possible to produce reliable estimates for even smaller datasets. Besides, other data sources may measure features not found in one dataset, which may give a richer picture of the relationship between different features. Ultimately, combining different data sources draws on the strengths and counterbalances the weaknesses of each source, resulting in more useful information, or lower costs, than what would be achievable from a single source [192].

However, leveraging data from other data sources can be challenging due to institutional differences with patients and the data coding and capture. As a result, the data from different sources may not deterministically be linked. [193] described four examples in which multiple data sources were combined to (1) extend and improve the coverage, (2) handle transitions from one approach of measurement of a variable to another, (3) correct errors in self-reported data, and (4) improve small-area prediction.

3.3 Techniques to Address the Challenges

3.3.1 Re-sampling Imbalanced Data

When the training data are extremely imbalanced, adjusting the class distributions (so that they are balanced) is currently the most popular approach to addressing the class imbalance problem. The two main sampling approaches include random oversampling (ROS) and random undersampling (RUS). Oversampling randomly duplicates the minority class samples, while undersampling randomly discards the majority class samples to modify the class distribution. Oversampling may cause overfitting as it replicates the same instances while undersampling may discard potential useful majority samples [186]. Chawla et al.

[152] introduced the Synthetic Minority Over-Sampling TEchnique (SMOTE), a sampling approach that over-samples the minority class (rare cases) by generating synthetic or artificial examples. These artificial data points are generated along the line segments between each minority data point and one of its k-nearest neighbors, thereby increasing the number of minority cases. Several hybrid methods have been proposed to improve SMOTE [155],[157],[158], [159].

SMOTE has been combined with ensemble methods in a variety of classification problems, including breast cancer classification [156],[157],[159],[160], [161],[162].

SMOTEBoost [156] combines SMOTE with boosting, which reweights the rare cases and compensates for skewed class distribution. RUSBoost [159] is a variation of SMOTEBoost, which combines boosting with random under-sampling (RUS). RUS removes examples randomly from the majority class until the desired balance is achieved. However, as stated above, these data balancing approaches may lead to correlated or less informative training data, thus producing suboptimal models, especially for severely imbalanced datasets as demonstrated in this study.

3.3.2 Transfer learning

In transfer learning, knowledge acquired from a source domain is transferred to a target domain, while the data distributions and feature spaces of the source and target may only partially overlap. Transfer learning can be used to solve the data dependence problem in areas with insufficient or unbalanced training data [194], [195], [196].

To mitigate the potential negative transfer of information from the source domain and to optimize the predictive performance on the target domain, the distributions of the target and source domain should be related [19],[20].

Transfer learning can be categorized into four categories: instance-based, feature-based,

network-based and adversarial-based transfer learning [20]. Instance-based transfer learning assumes that there are certain instances from the source domain that can be reused along with target data to improve the performance of target task [112]. Feature-based transfer learning builds a common feature representation, known as the latent space, to minimize the differences between the distributions of the source domain and the target domain while preserving the discriminability. In network-based transfer learning approach, network-based models are applied, and the transferred knowledge is encoded into parameters [197]. Finally, adversarial-based transfer learning methods learn new representations through adversarial learning process [194].

Notations Unless otherwise stated, the following notations will be used throughout this study.

Transfer learning can be defined as $\{D_s, D_t, f_t\}$, where $D_s = (X_s, Y_s)$ and $D_t = (X_t, Y_t)$ are source domain and target domain respectively, $X_s \in R^{N_s^0}$ is an N_s^0-dimensional feature space for the source and $X_t \in R^{N_t^0}$ is an N_t^0-dimensional feature space for the target, f_t is the predictive function for target domain. Theoretically, the objective of transfer learning is to help to improve the learning of the target predictive function f_t in D_t by transferring latent knowledge from D_s to D_t, where $D_s \neq D_t$. Suppose there are n_s and n_t observations for source data and target data respectively, typically, n_s is larger than n_t, and transfer learning uses this rich information to improve learning of the target predictive function f_t.

For generating generalizable transfer learning models, the key is to find out the relationship between the source and target domains, i.e., the latent transferable knowledge. For example, in deep neural networks, the latent transferable knowledge could be the latent space that minimizes the shifts between source and target. Such latent space can be generated by weights and biases of hidden layer $\mathbf{v}^k \in R^d$ which minimizes $\mathcal{D}(p(\mathbf{v}_s^l), p(\mathbf{v}_t^l)), 0 < l \leq L$. Here L is the total numbers of layers, and $\mathcal{D}$ denotes the distance measuring the difference

between the distribution of the source latent space $p(\mathbf{v}_s^l)$ and the distribution of the target latent space $p(\mathbf{v}_t^l)$, such as Kullback-Leibler (KL) divergence [198] and maximum mean discrepancy (MMD) [199]. We formulate the problem as follows. Suppose we observe data

$$\tilde{D}_s = \{(x_{s_i}, y_{s_i})|x_{s_i} \in X_s, y_{s_i} \in \{0,1\}\}_{i=1}^{n_s}, \tag{3.1}$$

$$\tilde{D}_t = \{(x_{t_i}, y_{t_i})|x_{t_i} \in X_t, y_{t_i} \in \{0,1\}\}_{i=1}^{n_t} \tag{3.2}$$

from both source domain and target domain. We assume that there exists a sequence of latent feature spaces $\{V_s^{k_j}\}_{j=1}^M$ and $\{V_t^{l_j}\}_{l=1}^M, k_j, l_j \in \{0,1,...,L\}$ for source and target task, which are spanned by the related hidden layers. The latent feature spaces satisfy (1) $d(V_t^{k_j}) = d(V_s^{l_j})$, with $d(\cdot)$ representing the dimension of the space, and (2) The distance between $V_t^{k_j}$ and $V_s^{l_j}$ is small enough, for $j = 1,...,M$. If the assumption does not hold, transfer learning may be unsuccessful, and may even result in inferior performing models in the target domain (i.e. negative transfer)[19]. Consequently, in this study, we transfer information from publicly available labeled healthcare datasets, which contain multiple discriminative features, to predict disease in an institutional dataset with potentially fewer discriminative features. Suppose we have the trained neural network $y = f(\mathbf{x})$ with L hidden layers, where $\mathbf{x}$ is the original feature and y is the response. Denote $\mathbf{v}_i^l = \zeta^k(x_i), l = 1,...,L$, to be l-th hidden layer related to the original feature x_i. Further denote $\mathbf{v}_i^r = \zeta^{k\to r}(\mathbf{v}_i^k)$, with $\zeta^{k\to r}$ represents the "transform" function between hidden layer $\mathbf{v}^k$ and $\mathbf{v}^r$. Consider the transfer learning from the source data $\tilde{D}_s$ to the target data $\tilde{D}_t$, the previous assumptions imply that the following conditions hold. There exist a sequence of transferable representations $\{\zeta_s^{k_j}\}_{j=1}^M$ and $\{\zeta_t^{l_j}\}_{j=1}^M$ for source data and target data respectively satisfying (1) $d\left(\zeta_s^{k_j}(x_{s_i})\right) = d\left(\zeta_t^{l_j}(x_{t_i})\right)$, and (2) $\mathcal{D}\left(p\big(\zeta_s^{k_j}(x_s)\big), p\big(\zeta_t^{l_j}(x_t)\big)\right) < \delta$, with δ to be certain threshold, for all $j = 1,...,M$. Once these two assumptions hold, the transfer learning can be conducted between $\tilde{D}_s$ and $\tilde{D}_t$ [200]. Our

goal is to build a neural network $y = f_t(\tilde{\mathbf{v}}_t^{l_m})$ based on the transferred latent space spanned by the transferred hidden layer $\tilde{\mathbf{v}}_t^{l_m}$, such that the target risk $r(f_t) = \|y_t - f_t(\tilde{\mathbf{v}}_t^{l_m})\|$ can be minimized. Here $\tilde{\mathbf{v}}_t^{l_m} = \zeta_s^{k_1 \to k_m}\left(\zeta_t^{l_1}(x_t)\right)$. As a result, the objective function for our transfer learning method is

$$\min_{f_t} \frac{1}{n_t} \sum_{i=1}^{n_t} \|y_{t_i} - f_t(\tilde{\mathbf{v}}_{t_i}^{l_m})\|, \tag{3.3}$$

where $m \leq M$ represents the number of layers to transfer.

Notice that the total number of layers suitable for transferring, which is denoted as M previously, is based on the predetermined threshold ϵ. A large ϵ means that more information can be borrowed from source data. However, if the borrowed information is too specific to the source data, it may results in an inefficient model to the target data. If ϵ is small, we turn to transfer latent features which are more similar between source and target. Then the transferred information could be too general. As a result, the threshold ϵ should be carefully chosen in transfer learning [200],[201],[202].

Deep transfer learning can be categorized into four categories: instance-based, mapping-based, network-based and adversarial-based deep transfer learning [20]

3.3.3 Deep Network-based Transfer Learning: Transfer learning with source models

With the recent dominance of deep neural networks (DNN), specifically Convolutional Neural Networks (CNN) for image classification, network-based transfer learning has become a standard training procedure, whereby a pre-trained network is used as a starting point for learning new tasks (images). Fine-tuning a network with transfer learning is usually much faster and easier than De Novo training a network with randomly initialized weights. The learned features can be quickly transferred to a new task using a smaller number of training images. Currently, a variety of popular pre-trained networks such as AlexNet [169],

GoogLeNet [203], VGGNet [204], and ResNet [205], have been trained on millions of images and can classify images into thousands of object categories. These pre-trained networks are publicly available and are used extensively by researchers for image classification through transfer learning. The primary motivations for such work are insufficient training data and the high computational cost involved in training a CNN from scratch. However, because transfer learning works best when the source and target tasks are related and images share underlying characteristics, application of transfer learning (especially pre-trained networks) to other areas such as structured EHR data can be more challenging. Here in case study 1, we employ a network-based transfer learning on Breast cancer to address a class imbalance in structured data.

3.3.4 Feature-based Transfer Learning: Transfer learning with unsupervised models

Feature-based transfer learning is another category of transfer learning. The interest of this approach is to learn a transformation to map the original data to a new feature space where the distance between different domains can be reduced implicitly or explicitly. We adopt autoencoders to construct a feature mapping from an original instance to a hidden representation. Autoencoders are unsupervised neural networks that learn a representation of data, not original data [206]. They use machine learning to do compression with backpropagation and broadly used in transfer learning with domain adaptation.

Here we mapped source and target data to joint distribution using autoencoder-decoder. We only extract the latent representation of both models and use these compressed features to classify the target data using a neural network classifier. Hidden layer added to smooth learning through backpropagation. Hyperparameters tuned to models best and drop-out used to prevent overfitting. Finally, binary cross-entropy with the sigmoid function used

for binary classification of output. Here in case study 2, we employ feature-based transfer learning to address poor performance learning on small dataset problems in structured data.

3.3.5 f-divergence

Given two points P and Q in a space S, we may define a divergence $D[P:Q]$ which measures their discrepancy. The standard distance is indeed such a measure. However, there are many other measures frequently used in many areas of applications [207].

The problem of f-divergence estimation is important in the fields of machine learning, information theory, and statistics. In probability theory, f-divergence is a function $D_f(P||Q)$ that measures the difference between two probability distributions P and Q. The divergence is intuitively an average, weighted by the function f, of the odds ratio given by P and Q. Many useful divergence measures belong to the set of f-divergences, independently introduced by Ali and Silvey [208], Csiszár [209], [209], and Morimoto [210] in the early sixties. KL-divergence and Hellinger distance are special cases of f-divergence, coinciding with a particular choice of f. In particular, for two probability distributions $p(x)$ and $q(x)$, one can define various measures of f divergence $D[p(x):q(x)]$ such as the Kullback-Leibler divergence and the Hellinger distance. A divergence is not necessarily symmetric, that is, the relation $D[P:Q] = D[Q:P]$ does not generally hold, nor does it satisfy the triangular inequality. It usually has the dimension of squared distance, and a Pythagorean-like relation holds in some cases. Divergence estimation is useful for empirically estimating the decay rates of error probabilities of hypothesis testing [211], extending machine learning algorithms to distributional features [212], and other applications such as text clustering [213].

In this study we use KL-divergence and Hellinger distance to quantify the difference between source and target dataset before and after encoding. Significant reduction in difference between two distributions prepares data for homogeneous transfer learning between source and

target.

Kullback-Leibler (KL) divergence To measure the difference between two probability distributions, a measure, called the Kullback-Leibler divergence, or simply, the KL divergence, has been popularly used in the data mining studies. The concept was originated in entropy and information theory.

The KL divergence, which is closely related to relative entropy, information divergence, and information for discrimination, is a non-symmetric measure of the difference between two probability distributions $P \in R^{k\in 1}$ and $Q \in R^{k\in 1}$. Specifically, the Kullback-Leibler (KL) divergence of Q from P, denoted $DKL(P,Q)$, is a measure of the information lost when Q is used to approximate P. $DKL(P,Q)$ is defined in following equation:

$$D_{KL}(p,q) = \sum_{i=1}^{n} p(x) log \frac{p(x)}{q(x)} \tag{3.4}$$

The KL divergence measures the expected number of extra bits required to code samples from P when using a code based on Q, rather than using a code based on P. Typically P represents the "true" distribution of data, observations, or a precisely calculated theoretical distribution [214],[215]. The measure Q typically represents a theory, model, description, or approximation of P.

Although the KL divergence measures the "distance" between two distributions, it is not a distance measure. This is because that the KL divergence is not a metric measure. It is not symmetric: the KL from P to Q is generally not the same as the KL from Q to P. Furthermore, it need not satisfy triangular inequality. Nevertheless, $DKL(P||Q)$ is a non-negative measure [216]. $DKL(P||Q) \geq 0$ and $DKL(P||Q) = 0$ if and only if $P = Q$.

Notice that attention should be paid when computing the KL divergence.

We know $\lim_{P\to 0} Plog_P = 0$. However, when $P \neq 0$ but $Q = 0$, $DKL(P||Q)$ is defined as ∞. This means that if one event i is possible ($i.e., P(i) > 0$), and the other predicts

it is absolutely impossible ($i.e., Q(i) = 0$), then the two distributions are absolutely different. However, in practice, two distributions P and Q are derived from observations and sample counting, that is, from frequency distributions. It is unreasonable to predict in the derived probability distribution that an event is completely impossible since we must take into account the possibility of unseen events.

Hellinger distance Hellinger distance[217] is a metric to measure the difference between two probability distributions. It is the probabilistic analog of Euclidean distance. Given two probability distributions, $P = (p_1, p_2, ..., p_k)$ and $Q = (q_1, q_2, .., q_k)$, Hellinger distance is defined as:

$$h(P, Q) = \frac{1}{\sqrt{2}} \cdot \sqrt{(\sum_{i=1}^{k} \sqrt{P_i} - \sqrt{Q_i})^2} \tag{3.5}$$

It is useful when quantifying the difference between two probability distributions. Hellinger distance satisfies triangle inequality. The reason for including $\sqrt{2}$ in the definition of Hellinger distance is to ensure that the distance value is always between 0 and 1. When comparing a pair of discrete probability distributions the Hellinger distance is preferred because P and Q are vectors of unit length as per Hellinger scale.

3.3.6 Comparing distributions

In our study, we apply the Kullback-Leibler (KL) divergence to measure the differences between the source and target domains. The KL divergence approach as relative entropy is a nonsymmetric measure of the divergence between two probability distributions [218], [219]. Such a measure has been widely accepted to define the "distance" between two distributions [220]. Given two probability distribution $p \in R^{k\times 1}$ and $q \in R^{k\times 1}$ defined on the probability space X, the KL divergence of q from p is the information lost when q is used to approximate p. We used the following formula to compute KL divergence between two datasets with a

different distribution.

$$D_{KL}(p,q) = \sum_{i=1}^{n} p(x) log \frac{p(x)}{q(x)} \tag{3.6}$$

The KL divergence can take on values in $[0, \infty]$. Particularly, if p and q are the exact same distribution ($p = q$) , then $D_K L(p||q) = 0$.

The other method to compare the discrepancy of two distributions is Maximum Mean Discrepancy (MMD). MMD is based on embedding probabilities in a reproducing kernel Hilbert space to find the discrepancy between two distributions. This measure has found numerous applications in machine learning and nonparametric testing. This distance is based on the notion of embedding probabilities in a reproducing kernel Hilbert space. In applications such as two-sample test and independence test that involve MMD, an estimate of MMD is constructed based on the estimates of μ_P and μ_Q respectively. The simple and most popular estimator of μ_P is the empirical estimator, $\mu_{Pn} := \frac{1}{n} \sum_{i=1}^{n} k(., X_i)$ which is a Monte Carlo approximation of μ_P based on random samples $(X_i)_{i=1}^{n}$ i=1 drawn i.i.d. from P [221].

Also, [222] proposed a nonparametric method to learn a piecewise constant function to approximate the underlying probability density function. Thier algorithm is defined based on the binary partition of σ and it uses Quasi Monte Carlo analysis to control the partition process.

As above mentioned there are several ways to compare the discrepancy of two distributions and here we utilized KL Divergence and Hellinger distance to measure the distance.

3.4 Related work

In this section, we briefly review related work in applying transfer learning, deep network-based transfer learning, and feature-based transfer learning in the healthcare area.

3.4.1 Transfer Learning in Structured Healthcare Data

Transfer learning methods have been recently applied in solving a wide range of real-world problems in healthcare applications. However, there are few studies of effectively using these methods in datasets other than images, audio, video, and raw text. Li et al. in their study of "constrained elastic net-based knowledge transfer for healthcare information exchange" proposed a model that can measure the differences among multivariate data distributions and based on this measurement they can avoid an unsuccessful transfer. They demonstrate the performance using the diabetes electronic health records (EHRs) which contains patient records from all fifty states in the United States. They successfully transfer the knowledge across different states to improve the diagnosis of diabetes in a certain state with insufficient records to build an individualized predictive model with the aid of information from other states [223].

Wiens et al. in their study of "A study in transfer learning: leveraging data from multiple hospitals to enhance hospital-specific predictions" applied similar transfer learning approaches to the medical data. They investigated an approach for building predictive models that involve augmenting data from individual hospitals with data from other hospitals. Using LIBLINEAR [224], they learned three different risk prediction models.
When data from two other hospitals are included in the training set, they saw a significant improvement in performance. These results demonstrate how auxiliary data can be used to augment hospital-specific models. Hospital A had only 82 positive training examples, compared with hospitals B and C with a combined 587 positive training examples. These additional positive examples helped the model generalize to new data. Comparing the performance of three classifiers, they saw that the classifier learned solely on data from the target task (i.e, Target-only) performs the worst [29].

Compared to our study, [225] provided a unified framework that potentially takes advantage

of auxiliary data using a transfer learning mechanism and simultaneously builds a robust classifier to tackle this imbalance issue in the presence of few training samples in a particular target domain of interest. They proposed a novel boosting-based instance transfer classifier with a label-dependent update mechanism that simultaneously compensates for class imbalance and incorporates samples from an auxiliary domain to improve the classification. They provided the details of the performance of various algorithms under different evaluation metrics using real-world datasets including Healthcare demographics, Parkinson dataset, and Text dataset. Their framework simultaneously compensated for the lack of data and the presence of class imbalance using a transfer learning model.

3.4.2 Deep Network-based Transfer Learning in Structured Healthcare Data

Similar transfer learning approaches have been applied successfully to medical data. [226] proposed a deep transfer learning framework using MeSH domain knowledge to improve automatic ICD-9 coding. First they trained a neural network for automatic MeSH indexing using BioASQ3 dataset. Then they used shared network (Convolutional Neural Network) and a fully connected layer with sigmoid function to predict the probability of the label on smaller target medical text. They demonstrated that transfer learning was the key component in improving automatic ICD-9 coding. Using transfer learning Micro-average F-measure, Micro-average Precision, Micro-average Recall increased from 0.39, 0.44, 0.35 to 0.41, 0.48 and 0.36. Their experimental results indicate that transfer learning is a key component to improve the performance of automatic ICD-9 coding and a deep learning approach is a foundation in the success of their proposed model. Compared to our approach, they did not try other transfer learning models to compare the performance of deep transfer learning model with a Feature based transfer learning model. [120] surveyed recent advances

in transfer and multitask learning for bioinformatics applications. Transfer and multitask learning offer an attractive alternative, by allowing useful knowledge to be extracted and transferred from data in auxiliary domains to help counter the lack of data problem in the target domain. Their survey shows that most current work on transfer learning is focused on sequence classification, gene expression data analysis, biological network reconstruction, biomedical text, image mining, and sensor-based ubiquitous healthcare. Similar to our paper, their study is on transfer learning in biomedical application, however their research review is more comprehensive including research work in transfer learning and multitask learning in the area of bioinformatics and biomedical applications.

Gupta et al. in their study of "Transfer Learning for Clinical Time Series Analysis Using Recurrent Neural Networks" investigate that to what extent transfer learning can address issues of training deep RNNs in clinical time series. Training deep Recurrent Neural Networks (RNNs) requires such a large labeled data, high computational resources, and significant hyperparameter tuning effort in clinical time series. Authors demonstrate that (i) models trained on features extracted using RNN source model outperform or, in the worst case, perform as well as task-specific RNNs; (ii) the models using features from source models are more robust to the size of labeled data than task-specific RNNs; and (iii) features extracted using source RNN model are generic enough and perform better than typical statistical hand-crafted features [227]. Using transfer learning the performance of classification improved from 0.90 to 0.95. Similar to our study they used neural network based model to implement transfer learning, however they used Logistics regression at the end of the layers for classification.

3.4.3 Feature-based Transfer Learning in Structured Healthcare Data

Unsupervised feature transfer enables neural networks to transfer latent layer features for a classifier trained in a supervised way. Chetak Kandaswamy et al. in their paper [200] proposed training stacked denoising autoencoders (SDAs) on a source problem and transferring its features to help to solve the target problem. Similar to our approach, they implemented both feature based transfer learning and network-based transfer learning on their study. For example they pick the features of the model built to classify images of digit from 0-to-9 and reused them to classify images of letters from a-to-z. Similar experiments were conducted by reversing the role played by each problem. For unsupervised transfer: First, they transferred unsupervised features of stacked denoising autoencoders from source to target problem. Then they fine-tuned the entire multi-layer perceptron with back-propagation on the target problem. For supervised transfer: they transferred pre-trained k layers from source to target and then they fine tuned the whole target model. In contrast to our method, they used Logistic regression for classification. They showed that SDAs using the unsupervised feature transfer outperform regular models. They achieved 7% relative improvement on average error rate and in the case of supervised feature transfer they achieved 5.7% relative improvement in the average error rate.

[228] stated a successful transfer using stacked denoising auto-encoders arises in the context of domain adaptation, i.e., where one trains an unsupervised representation based on examples from a set of domains but a classifier is then trained from few examples of only one domain. Their paper focused on the context of the Unsupervised and Transfer Learning Challenge where they care about predictions on examples that are not from the same distribution as the training distribution. In their paper, they refer to related work in deep transfer learning explaining that deep learning seems well suited to transfer learning because

it focuses on learning representations and in particular "abstract" representations, representations that ideally disentangle the factors of variation present in the input. As opposed to our paper, they have not reported any case study of actual implementation of different types of transfer learning.

Similar to our method on feature based transfer learning model, [229] trained a neural network to predict source task with present labels that were not directly related to phenotype targets but contained enough information to provide a training signal for learning a useful representation of raw EHR data. They specifically applied a variation of the split-brain autoencoders (med2rx and rx2med models) with two hidden layers to predict the prescription from diagnostic codes. Then they used these auto-encoders as fixed features extractors in the logistic regression model trained to predict the target task using a much smaller number of reliable samples. Using auto-encoders as feature extractors improved AUC from 0.74 on baseline to 0.84 in predicting Essential Hypertension. AUC improved from 0.76 to 0.79 for predicting Diabetes in patients using transfer learning. Unlike our paper, they have not reported other performance measurements such as accuracy, F-score, recall using feature based transfer learning in order to generalize the improvement in performance of feature based transfer learning model.

3.5 Deep Transfer Learning on Structured Data

The Deep Neural Network Architecture To better understand the implementation of the deep transfer learning (DTL) in this study, we briefly review the formulation of the DNN, which is simply the conventional multilayer perceptron with more than two hidden layers [230]. Figure 3.1 depicts a DNN with $L+1$ layers: an input layer, $L-1$ hidden layers, and an output layer. For simplicity, we call the input layer as layer 0, and the output layer as layer L, thus the hidden layers are $l = 1, ..., L-1$. Given a data point $\mathbf{x}$ (patient features), its

corresponding binary label or output unit $y \in \{0, 1\}$ (benign/malignant breast cancer and diabetic/non-diabetic) can be obtained in DNN through a series of mappings of weighted sums of the inputs over the hidden layers as follows

$$\mathbf{v}^0 = \mathbf{x} \tag{3.7}$$

$$\mathbf{v}^l = \eta(\mathbf{z}^l) = \eta(W^l \mathbf{v}^{l-1} + \mathbf{b}^l), \quad 0 < l \leq L \tag{3.8}$$

where for layer l, $\mathbf{z}^l = W^l \mathbf{v}^{l-1} + \mathbf{b}^l$ is called the excitation vector, $\mathbf{v}^l$ is the activation vector, W^l is the weight matrix, and $\mathbf{b}^l$ is the bias vector.

Each layer has a specified number of neurons N^l and $\eta : R^{N^l} \rightarrow R^{N^l}$ is the activation function applied to the excitation vector. The output layer $l = L$ represents a binary classification task; so we use a sigmoid function $\eta^l(\cdot)$ to produce class probabilities

$$\eta^l(x) = \frac{1}{1 + e^{-x}} \tag{3.9}$$

The threshold is applied to the cut-off point in class probabilities, which is 0.50 to assign class labels. By evaluating the true positive and false positives for different threshold values, a curve can be constructed that stretches from the bottom left to top right and bows toward the top left which presents the ROC curve. For the activation function η we used the exponential linear unit (ELU) [231], which is known to speed up learning in DNN and can improve classification accuracy. The ELU is defined by

$$\eta(x) = \begin{cases} \alpha e^x - \alpha, & \text{if } x \leq 0 \\ x, & \text{if } x > 0 \end{cases} \tag{3.10}$$

where hyperparameter α controls the value to which the ELU saturates for negative inputs. We used the default value 1 in the deep learning software package [232]. By minimizing the cross-entropy loss function, the model parameters $\{W^l, \mathbf{b}^l\}$, $0 < l \leq L$ were learned through the backpropagation algorithm. We used the grid search capability from the scikit-learn python machine learning library to tune the hyperparameters of deep learning and transfer learning models.

Autoencoder-decoders Unsupervised Learning deals with data without labels. An example of Unsupervised Learning is dimensionality reduction, where we condense the data into fewer features while retaining as much information as possible. An auto-encoder uses a neural network for dimensionality reduction. This neural network has a bottleneck layer, which corresponds to the compressed vector. When we train this neural network, the 'label' of our output is our original input. Thus, the loss function we minimize corresponds to how poorly the original data is reconstructed from the compressed vector.
Autoencoder-decoder can be represented two segments of neural network.

$$\phi : X \rightarrow F \tag{3.11}$$

$$\delta : F \rightarrow X \tag{3.12}$$

$$\phi, \delta = argmin||X - (\phi \cdot \delta) \times X||^2 \tag{3.13}$$

The encoder function, denoted by ϕ, maps the original data X, to a latent space F, which is present at the bottleneck. Bottleneck contains the compressed representation of the input data. This is the lowest possible dimensions of the input data. The decoder function, denoted by δ, maps the latent space F at the bottleneck to the output. The output, in this case, is the same as the input function and Reconstruction Loss measures measure how well

the decoder is performing and how close the output is to the original input. Thus, we are basically trying to recreate the original data after some generalized non-linear compression. The training then involves using back propagation in order to minimize the network's reconstruction loss.

We implemented two types of DTL on publicly available diabetes dataset: Network-based transfer learning and feature based transfer learning. In Network-based transfer learning, we trained source model and saved weights and biases of the first and second hidden layers. Then we transferred the saved network to the target model with the same architecture as the source model and retrained the target model. On the other hand, in feature based transfer learning, autoencoders aroused in the context of domain adaptation where unsupervised representation trained from a set of domains but the classifier was then trained only from the target domain.

For that purpose, bottleneck layers of source and target were saved and transferred to the target DNN model. This bottleneck forces a compressed knowledge representation of the original input. If the input features were each independent of one another, this compression and subsequent reconstruction would be a very difficult task. However, if some sort of structure exists in the data (ie. correlations between input features), this structure can be learned and consequently leveraged when forcing the input through the networks bottleneck. In the representation-learning phase of autoencoder, some of the elements that explain marginal distribution for Y in the training classes are captured by the learned representation. The output of the encoder is a set of neurons that forms the encoding (compressed set of features), so the learner trained on the target set just needs to pick up those elements relevant to the discrimination among target set classes.

3.6 Baseline Methods for Imbalanced Classification

For performance comparison, we trained four popular baseline machine learning methods: logistic regression (LR), random forest (RF), DNN, and stochastic gradient boosting (XGBoost) on the target data. We experimented with 3 and 5 hidden layers in DNN. XGBoost [233] is an ensemble classifier which works on the principle of gradient boosting decision trees, where decision trees are sequentially added in the model, and each tree corrects the errors of the previous tree in the sequence. It therefore produces a strong model from a collection of weak models.

To address the imbalanced learning problem, we used SMOTE as a preprocessing step to oversample the data and create more balanced data for training the baseline classifiers. We used k = 5 nearest neighbors in the SMOTE algorithm. SMOTEBoost and RUSBoost were also included in the baseline models.

3.7 Training, Validation, and Evaluation

We implemented the DNN models using the Keras Python library [232], wherein the Patience hyperparameter (Table 3.4) is set to monitor the performance of the model and trigger early stopping. A 5-fold cross validation procedure was used, wherein one fold was used for validation, one fold used for testing, and the rest used for training the models. The models were evaluated using AUC, F1- measure (F1-Score), and sensitivity (Sens). The software code for the DTL model implemented in this study can be downloaded from GitHub [2].

[2] `https://github.com/AydaFarhadi/TransferLearning`

3.8 Case Study 1

3.8.1 Data

Source Data

Breast Cancer Datasets We used two publicly available breast cancer datasets from the UCI machine learning repository [172] as source data to implement the proposed DTL on structured data. The breast cancer datasets include: (1) The Wisconsin Diagnostic Breast Cancer (WDBC) with 357 benign and 212 malignant cases.

Features have been computed from a digitized image of a fine needle aspirate (FNA) of a breast mass. Breast Cancer appears as a result of mutations, or uncommon evolve in the genes responsible for regulating the growth of cells and keeping them normal. The genes are in each cell nucleus, which acts as the "control room" of each cell [234]. In order to capture the features contributing in breast cancer, ten initial real-valued features were computed for each cell nucleus such as radius (mean of distances from the center to points on the perimeter), texture (standard deviation of gray-scale values), perimeter, area; smoothness (local variation in radius lengths), compactness ($perimeter^2/area - 1$), concavity (severity of concave portions of the contour), concave points (number of concave portions of the contour), symmetry, and fractal dimension. The mean, standard error, and mean of the three largest values of these features were calculated for each image, resulting in 30 features (2) The Wisconsin Prognostic Breast Cancer (WPBC) data represents follow-up data for patients seen at the University of Wisconsin Hospital at Madison, Wisconsin from 1984 until 1995. It included only cases with invasive breast cancer and no evidence of distant metastases at the time of diagnosis. As with the WDBC data set, 30 of the 34 features in the data were calculated for each image. There were 47 breast cancer recurrence and 151 non-recurrence cases in WPBC.

The WDBC and WPBC contain 30 similar features, therefore we merge the two datasets vertically by these features to create a single source dataset to generate the source model. We have overall 767 in the merged dataset (Details of features are presented in Appendix, table A.2).

Target Data

The Mayo Mammography Health Study (MMHS) The primary data used to demonstrate the effectiveness of the proposed transfer learning method on structured data consists of 15,386 women, $age \geq 37$, diagnosed with breast cancer between January 15, 2009, and January 15, 2016. This data is a subset from the Mayo Mammography Health Study (MMHS) Cohort, a cohort of 19,936 women enrolled at the Mayo Clinic, Rochester, with a screening mammogram performed between 2003 and 2006. Only women in the full cohort who had at least one full-field digital mammograms performed during the follow-up period were eligible for this study. Information was collected through the EMR, self-reported and clinical questionnaires.

The Mayo Clinic institutional review board approved the MMHS. The dataset for this study contains patients demographics, body mass index (BMI), family history, and serial measurements of clinical breast density, assessed according to the 4-category Breast Imaging Reporting and Data System (BI-RADS). Table 3.1 presents the distribution of features selected for this study stratified by diagnosis status. During the follow-up period, 487 (3.2%) women were diagnosed with invasive breast cancer (malignant), significantly underrepresented compared to those without breast cancer (benign), 14,899 (96.8%). Thus, the MMHS data resulted in severely imbalanced data (IR = 30.6). This, in addition to the fact that only a limited number of features was used for these women (see Table 3.1), made the data challenging for standard classification methods to learn. Further details about the MMHS can be found in [172],[235] .

To handle missing values in the data, we employed two imputation techniques. First, we used LOCF(Last Observation Carried Forward) and NOCB (Next Observation Carried Backward) [236] to impute the serial breast density levels. Specifically, missing levels in breast density measurements were imputed by carrying forward the last most recent valid non-missing level. We then converted the serial values from a long format to a wide format, such that the individual values at each time point became feature vectors. We then applied MissForest to impute all other missing features.

UCI Mammographic Mass Data This is a publicly available dataset collected at the Institute of Radiology of the University Erlangen-Nuremberg between 2003 and 2006 [237] and available from the UCI machine learning repository [175]. The data was used in this study to predict the classification (benign or malignant) of a mammographic mass lesion. It contains 516 benign and 445 malignant cases. Because the mammographic mass dataset was relatively balanced (IR = 1.2), the baseline classifiers were expected to perform well. As such, we also used the data to investigate the effect of imbalanced learning by training the models under different IRs in the training data. Specifically, we carried out a controlled experiment by simulating class imbalance in the training data, whereby samples from the malignant class were removed randomly. We created training datasets where the proportion of observations in the minority class was: 10% (IR = 5.4), 5% (IR = 10.7), and 2% (IR = 26.8).

We imputed missing values in the mammographic mass data set with MissForest.

3.8.2 Method

Proposed Deep Transfer Learning for Breast Cancer

The proposed deep transfer learning is based on the DNN described above; where the goal is to transfer knowledge from a structured source domain data with potentially large and

balanced class distribution to a structured target domain data that is imbalanced (see table 3.3). We use "TensorFlow" deep learning library in python to develop our neural network models [238].

The model consists of an input layer, an output layer and three hidden layers with a total of 8334 parameters.

The weights were initialized with random values and the backpropagation algorithm was used to optimize the weights. The hyperparameters of the Breast Cancer Network were tuned using grid search and presented in Table 3.4. By the theory of DNN, the initial layers (say $l = 1, 2$) capture generic features about the disease (cancer), while the later layers focus on specific disease characteristics (e.g. breast cancer pathology). Therefore, given a source DNN model generated on one dataset, we can freeze (fix the weights) one or more initial layers, and retrain the remaining layers on another dataset. Thus, in DTL, we essentially use the same DNN architecture, where the initial layers of the source model are used as a starting point for retraining on the new data. Note, however, that in transfer learning, any of the layers of the DNN may be frozen (weights are not updated after each training epoch) or unfrozen (weights are updated after each training epoch), and as such these layers are referred to as transferred hidden layers.

we also develop our target DNN using 5 layered DNN and 7 layered DNN based on the Mayo Clinic Breast cancer data. We implemented both freeze transferred hidden layers and unfreeze transferred layers, with a different number of layers. Specifically, we implemented and compared the performance of the DTL models represented on Table 3.5.

Figure 3.2 illustrates the implemented DTL network architecture, where the source model (shown in blue) is a DNN. To derive the 5L-2T-Freeze model, for example, we transferred the top two hidden layers of the source model to target model and froze their weights [239].

To generate the source DNN model, we set the learning rate to 0.001 and 0.0001 for the retraining model. Setting a smaller learning rate for the retraining model ensures that large

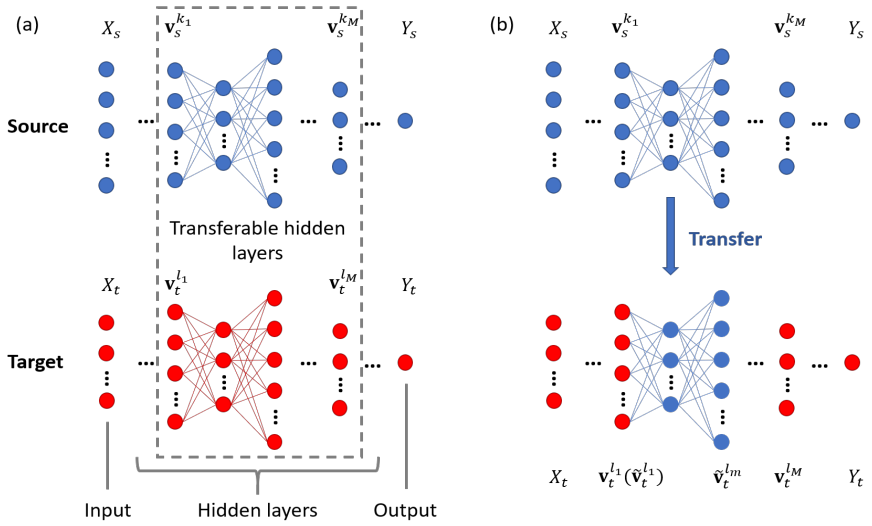

Figure 3.1: Illustration of transfer learning. Left: locating transferable hidden layers. Right: transferring layers that can improve the learning of the target predictive function.

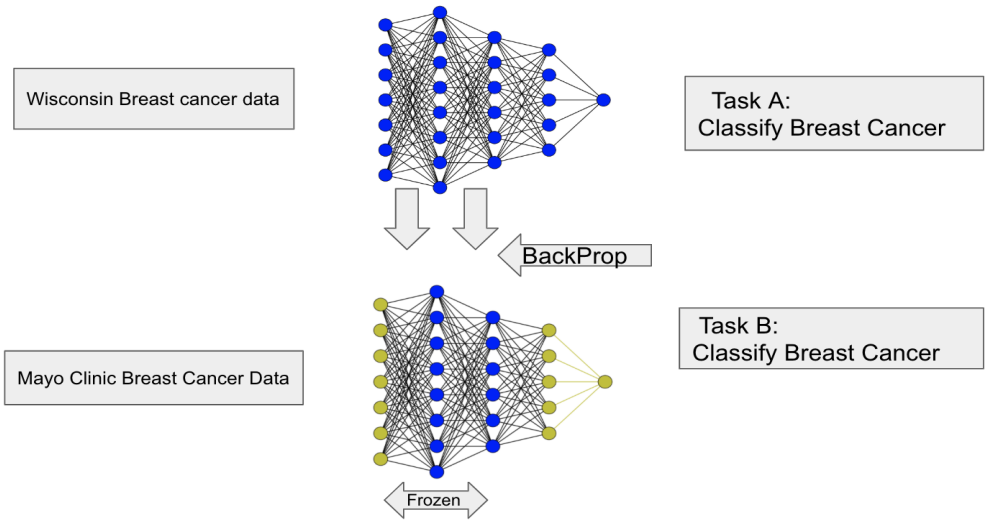

Figure 3.2: Deep transfer learning for breast cancer classification

gradient updates do not distort the pre-trained weights [240],[241].

3.8.3 Results

The performance of classifiers on Breast Cancer dataset Figure A.2 in Appendix presents mean of the 10 original features in combined WDBC and WPBC source data with a total of 767 observations. The minority class was composed of cases with malignant tumors (from WDBC) or recurrent cancer (from WPBC) with an IR of 1.96. Thus the source dataset was relatively balanced.

Basic descriptive statistics of the features in the MMHS target breast cancer data stratified by breast cancer diagnosis is presented in Table 3.1. The table includes 6 BI-RADS breast density assessments over six time periods (years). Each of the 4 BI-RADS levels represents gradations of the likelihood that a cancer exists, from lowest to highest probability. We also include a transformed variable with levels (0, -1, and 1) indicating no change, decrease, and increase in the BI-RADS respectively over the six time periods.

Performance of Source DNN on Source Data

We used AUC to select the best model parameters during training and validation in predicting breast cancer on the source data. The corresponding validation AUC on breast cancer source model was 0.96(0.10).

Performace of DTL and Baseline Models

We first trained the baseline models on the target datasets without sampling, and then retrained the models on oversampled data using SMOTE. For oversampling, only the training portion of the cross-validation was sampled. The validation and test set were left unchanged. We equally trained the DTL models: 5L-2T-Freeze, 7L-2T-Freeze, 5L-2T-UNFreeze, and 5L-

1T-UNFreeze as previously described to classify breast cancer on the MMHS data and the UCI Mammographic mass datasets.

Table 3.10 presents performance results of the models on the MMHS data. With respect to the AUC metric, resampling the data with SMOTE marginally improved the performance of LR, while the performance of RF deteriorated. As previously said, the poor performance of RF on the oversample data was likely due to the introduction of uninformative or correlated examples by SMOTE. The performance of RUSBoost, which creates a balanced training set by undersampling rather than generating synthetic examples, was significantly better than all baseline models including SMOTEBoost. On the other hand, all the DTL models except 7L-2T-Freeze outperformed the other comparator methods. The 5L-2T-UNFreeze showed the best AUC performance. However, the 7L-2T-Freeze model was the most sensitive model in detecting malignant cases.

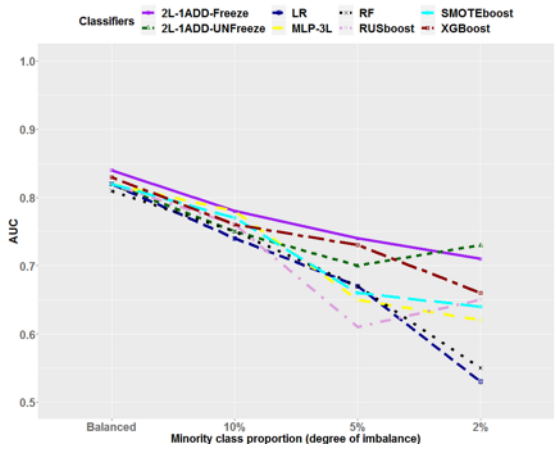

Figure 3.3: Performance of models on the UCI Mammographic mass data with simulated imbalanced training data distribution

Figure 3.3 presents performance results of the models on the UCI Mammographic Mass data. As expected, the performance of the models deteriorated as the proportion of minority classes in the data decreased (or increased in IR). The AUC values for Balanced in the figure

represents the performance of the models on the original dataset (IR = 1.2). The 5L-2T-Freeze (2L-1ADDFreeze) DTL model showed slightly better performance compared to other methods, and remained dominant throughout the different simulated levels of IRs. Finally, the 5L-2T-Freeze and 5L-2T-UNFreeze (2L-1ADD-UNFreeze) DTL models significantly outperformed all the other models for the very severe imbalanced (IR = 26.8) case. We omitted the results for the 7L-2T-Freeze and 5L-1T-UNFreeze DTL models as they slightly underperformed compared to the 5L-2T-Freeze and 5L-2T-UNFreeze but performed better than most of the baseline models.

Interestingly, the performance of some of the methods actually increases from a 5% minority class rate to the 2% rate. It is unclear what may be the reason for this increase. A contributing factor may be due to the randomness in which we simulated the imbalanced class levels. Based on the comparison results, it can be concluded that the DTL is an efficient methodology for standard classification tasks with balanced or imbalanced structured healthcare data.

3.9 Discussion

Breast cancer is a cause of significant morbidity and mortality in the US and worldwide. Early detection of high grade malignant breast cancers is critical for improved survival and other patient outcomes. Yet, despite the large number of women with the disease, effective early identification of malignant vs. benign lesions with digital mammography (the primary modality of breast cancer screening) is difficult due to the small numbers in the overall population. Traditional methods for predicting rare events do not provide adequate accuracy. In this work, we demonstrated that transfer learning on structured data is an efficient learning technique that can mitigate the class imbalance problem prevalent in most disease classification tasks in healthcare.

In healthcare, correctly detecting the rare events or minority class is crucial, as they correspond to high-impact events and misclassifying them can carry significant consequences. For example, misclassification of non-cancerous cells as malignant (i.e. false positive) may lead to unnecessary clinical testing (with its own side effects and harms) and patient anxiety. Similarly, misclassification of cancerous cells as benign (i.e. false negative) results in delayed diagnosis, higher burden of disease at the time of detection, and ultimately lower probability of successful treatment and survival [242], [171].

This work proposes an accurate and efficient deep transfer learning approach (DTL) for improving early detection of breast cancer as a test case for other similar problems in health care by addressing the key issue of imbalanced healthcare/medical data. DTL has commonly been used to deal with lack of data when the input is non-structured (e.g. images), however, its application to structured data to address the class imbalanced problem has not been investigated. We demonstrated in this work that DTL from a structured and relatively balanced source data can be used to facilitate accurate modeling in a more general, imbalanced problem. We illustrated our claims in a real healthcare problem by facilitating the prediction of breast cancer occurrence in a general screening cohort (MMHS data) by transferring features learned from differentiating between malignant and benign or recurrent and non-recurrent cancers (public WDBC and WPBC data).

The results reflect known research for imbalanced data; as a dataset skews further towards extreme imbalance, predictive algorithms tend to perform worse (at times significantly so). For once, the signals which may be predictive of the minority class shrinks with respect to the noise. Further, the loss function of the learning algorithm no longer accurately reflects the intended problem, but naturally prioritizes learning the majority cases. As with many problems, attempts to mitigate these problems in our use-case using resampling or reweighting show only minimal benefit, or in some cases disadvantageous. We hypothesize that the

real features which are needed to learn to differentiate the classes (at risk for breast cancer and not at risk) is not strongly present, which prevents the classifiers from benefiting from data resampling or reweighting. In other words, the models based on resampled data alone are not sufficiently trained. In contrast, transfer learning can be leveraged to learn these features without significant manipulation of the target data.

In traditional transfer learning application problem, the source data is typically very large, as it is assumed that large dataset is needed to develop a well-trained model. However, we show that despite this common assumption, the features learned are not completely invalidated. In fact, the features actually benefit a new task where there was previously a lack of data. We believe this is due to the source data being rich in discriminative information compared to the target data. This finding is potentially extremely useful as many clinical decision support tools have poor accuracy due to too little signal in the training cohort.

Limitations of this study should be noted. First, the performance of this method was evaluated only on breast cancer datasets, which limits the conclusions that can be drawn from this experiment. However, the results are encouraging and we plan to apply this method in other real-world healthcare datasets. Second, real-world healthcare data is often afflicted with high levels of missing data for various reasons. Although we used a model-based method to address missing values, it is unclear how this may have biased our model. In the future, we plan on evaluating the method on datasets with varying degrees of missingness. Third, we implemented a simple technique for simulating imbalanced data by randomly dropping minority examples. A more systematic approach that controls for the imbalanced degree can be implemented [189].

3.10 Case Study 2

3.10.1 Data

Source Data

Diabetes datasets We also used the dataset representing 10 years (1999-2008) of clinical care for diabetes patients at 130 US hospitals and integrated delivery networks, from the University of California Irvine, to generalize our models of transfer learning on Structured dataset. The dataset encounters satisfy the following conditions:

- It is an inpatient encounter (a hospital admission)
- It is a diabetic encounter, that is, one during which any kind of diabetes was entered into the system as a diagnosis.
- The length of stay was at least 1 day and at most 14 days.
- Laboratory tests were performed during the encounter.
- Medications were administered during the encounter.

The target variable of readmission had two levels of '< 30' (patient readmitted within 30 days), '> 30' (patient was readmitted after 30 days) which we defined as 'yes' and 'no' (patient was not readmitted). We defined it as a binary classification problem. It includes over 49 features representing patient and hospital outcomes. The Total number of instance are 100,000 data [243].

Table 3.7 indicates features of diabetics readmission dataset. Missing values in the data were imputed using the MissForest [244] algorithm. MissForest iteratively trains a random forest model on the observed (non-missing) portion of the data to predict the missing values.

Target Data

Pima Indians Diabetes Dataset This is a publicly available dataset from the UCI machine learning repository [175]. The dataset includes the following features: number of times pregnant, Plasma glucose concentration 2 hours in an oral glucose tolerance test, Diastolic blood pressure (mm Hg), Triceps skinfold thickness (mm), 2-Hour serum insulin (mu U/ml), Body mass index (weight(kg)/$(height(m))^2$), Diabetes pedigree function and Age (years). A total of 768 cases are available in the Pima Indians diabetes dataset.

3.10.2 Method

Proposed Deep Transfer Learning for Diabetes

We implemented two types of transfer learning on this dataset. Mapping based transfer Learning and Network-based transfer Learning. To implement mapping-based transfer learning: First we developed two autoencoder-decoders for the publicly available Diabetes readmission dataset and Pima Indian diabetes dataset respectively (Fig.3.4, Fig.3.5). Each network has an encoder, bottleneck layer and a decoder. We saved the bottleneck layer of both models in order to transfer them to the target model for classification. Proposed Deep Transfer Learning for Diabetes using Mapping based transfer learning is presented here. The

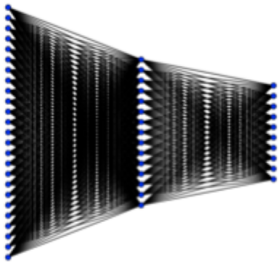

Figure 3.4: Source encoder

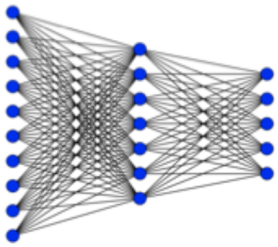

Figure 3.5: Target encoder

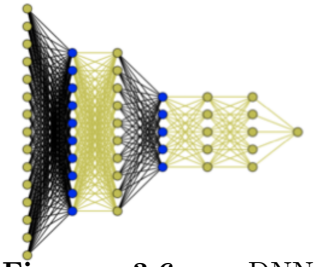

Figure 3.6: DNN model

DNN model was similar to the one described for Breast Cancer with differences in the optimized hyperparameters (see table 3.4). We developed DNN using a 5 layered DNN based

on the Pima Indian diabetes dataset described above.

Also we developed network-based transfer learning similar to Breast cancer dataset with different number of hidden layers and optimized hyperparameters (see table 3.8, table 3.9). The goal was to transfer knowledge from structured source domain data with a large amount of data to a structured target domain data with a smaller amount of data.

Table 3.8 shows the parameters of the model with details of l- layers, W - weights and b - biases. The weights were initialized with random values and the backpropagation algorithm was used to optimize the weights. The hyperparameters of the network were tuned using grid search and presented in Table 3.9. Our optimization problem was reducing the distance between encoded representation and target dataset using backpropagation on DTL classifier.

We used max-val-ACC to save the model only when there is an improvement in ACC. Batch normalization used after encoded layer to reduce the internal covariate shift by preventing outliers in the aggregate code distribution. Given a source DNN model generated on one dataset, we can freeze (fix the weights) one or more initial layers, and retrain the remaining layers on another dataset. Therefore, in DTL, we essentially use the same DNN architecture, where the initial layers of the source model are used as a starting point for retraining on the new data.

3.10.3 Results

Divergence between source and target features in Diabetes dataset As a good and widely accepted measure of the distance between two distributions, KL divergence between source and target data is applied to describe the difference between the source and target domains. We compare KL divergence before and after encoding to check the possibility of re-

ducing KL divergence between source and target data. Difference in dimensions of source and target dataset made calculating KL divergence challenging. We applied Principle Component Analysis (PCA) on original data to remove redundancy in the dataset. PCA is a linear transformation with a well defined inverse transform [245]. On the other hand, autoencoders can have nonlinear encoder/decoders. As a result, we extracted Principle Components (PCs) and encoded features on two datasets. Then we compared KL divergence between PCs of source and target data, and encoded features of source and target data using a method as explained below.

Let KL_{before} denote the KL-divergence before encoding, and KL_{after} denote the KL-divergence after encoding. Consider the following hypothesis:

$H_0 : KL_{after} = KL_{before}$.
$H_1 : KL_{after} < KL_{before}$.

For such hypothesis testing, we first randomly chose 200 samples from the source PCs and target PCs, then calculated the KL-divergence. In particular, the densities for the samples can be estimated using Kernel density estimation. With the estimated densities for source PCs and target PCs, the KL-divergence can be obtained. There are some existing methods in calculating the KL-divergence, here we use the R package Fast Nearest Neighbor Search Algorithm (FNN) [246], [247], [248]. In this package, it first estimates the distributions p and q for source data (dimension: n1*d, in the paper, is 200 * 10) and target(dimension: n2 * d, in the paper, is 200 * 10) using K nearest neighbors such that for small regions it needs to search small area and for large region it needs to search in large distances to find nearest neighbors. Then it calculates the KL-divergence by summing up the divergences of $q(x)$ from $p(x)$ and $p(x)$ from $q(x)$.

By repeating this for 300 times, we could get the empirical distribution of the KL-divergence between source PCs and target PCs (see the red curve in Figure. 3.7). Then we conducted

the same experiment for encoded features and got the empirical distribution of the KL-divergence between the encoded features (marked as the blue curve in Figure 3.7). However, the work on this is preliminary and is undergoing further validation for a journal submission. Based on the empirical distributions shown in Figure 3.7, we conducted the two-sample test. The p-value for the test was nearly 0 (p-value $< 2.2 \times 10^{-16}$). The testing result confirmed that implementing the Autoencoder-decoder could reduce the KL-divergence between the source and the target, and result in proper domain adaptation for transfer learning [249]. We applied the feature based transfer learning method (3.3) to this data and chose δ to be $D_{KL}(x_s, x_t)/8$.

Next we compared Hellinger distance between PCs of source and target data, and encoded features of source and target data. We computed the Hellinger distance between features of two datasets and then we took an average of all divergences. Assume the δ is Hellinger distance between between PCs of source and target dataset. Encoding prepared data for transfer learning by reducing δ to $D_{HD}(x_s, x_t)/6$ between source and target dataset (Difference reduced from 0.043 to 0.07).

The performance of classifiers on Diabetes dataset

Performace of Source DNN on Source Data

We used ACC to select the best model parameters during training and validation in predicting diabetes on source data. The corresponding ACC on diabetes source model was 0.91(0.03).

Performance of Source DNN on Source Data

We used ACC to select the best model parameters during training and validation in predicting diabetes on source data. The corresponding ACC on diabetes source model was

0.91(0.03).

We first trained the baseline models on the target datasets. Then, we equally trained the DTL models: 6L-2T-Freeze, 7L-2T-Freeze, 6L-2T-UNFreeze, and 7L-2T-UNFreeze as previously described to classify breast cancer on the Pima Indians Diabetes dataset. Finally, we trained the feature based DTL model.

Figure 3.8 shows the model loss before transfer learning against the training epochs of the target data. Figure 3.9 shows the model loss after transfer learning against training epochs of the target DNN in predicting diabetes on the target data. Before transfer, training loss decreases rapidly throughout the training epochs, while the validation loss initially decreases rapidly to 0.45, levels off on 0.41 with high variation, and then start increasing slowly at around 60 epochs. However, after transfer, both training and validation loss decreases slowly to 0.40 with less variation throughout the training epochs, level off at 0.38, and then starts increasing slowly at around 80 epochs. Consequently, after the transfer, 80 was considered as the converge point; the point at which the model test performance starts to deteriorate, thus indicating overfitting is in effect.

Performance of DTL and Baseline Models

The performance of classifiers are compared using python for classification. Using source model on diabetes readmission dataset improved the accuracy on target data to 0.85 using DTL-5L-Feature (Features extracted via unsupervised learning to target NN model to retrain source network on the target model).

Table 3.10 presents performance results of the models on the UCI Pima Indians diabetes data. As expected, with cross-validation we get a true performance of the models. Feature based DTL model with 5L outperforms all other models of classification. On the second level, the 6L-2T-UNfreeze DTL model outperforms baseline models of classification.

3.11 Discussion

Diabetes means blood sugar is above the desired level on a sustained basis and it is one of the most wide-spread diseases. Diabetes can cause serious health complications including blindness, blood pressure, heart disease, kidney disease and nerve damage, etc. which is hazardous to health. Early detection of this disease could significantly decrease healthcare costs via the diagnosis of diabetes. Yet, despite a large number of researches in this area, medical studies demonstrated that diabetes pathology is increasing in the last decades and the trend does not tend to stop [250]. Traditional methods for predicting disease in small datasets do not provide adequate accuracy. In this work, we demonstrated that transfer learning on structured data is an efficient learning technique that can make efficient classification when a large amount of labeled datasets in healthcare is not accessible.

We demonstrated in this work that DTL from a structured and large balanced source data can be used to facilitate accurate modeling in a more general problem. In the unsupervised representation-learning phase of autoencoder, some of the items that explain $P(X|Y)$ for Y in the training classes are captured by the learned representation and they are useful in predicting different classes from the target data set. Among different approaches to transfer learning, the feature-based transfer learning methods have proven to be superior for the scenarios where original raw data between domains are very different while the divergence between domains can be reduced. Encoders discover features that capture the generic items of variation present in all the classes, so the classifier trained on the target set just needs to pick up those items relevant to the discrimination among target set classes. We illustrated our claims in real healthcare problems by facilitating the prediction of diabetes occurrence by transferring features learned from differentiating between diabetic and non-diabetic patients.

Limitations of this study are similar to the ones listed for Case study 1.

3.12 Conclusion

Healthcare is one of the most common domains suffering from scarcity of labeled data and imbalanced data, and addressing them are an urgent priority.

Unlike traditional machine learning methods, in which the creator of the model has to choose and encode features ahead of time, deep learning enables a model to automatically learn features that matter. In this way, a deep learning model learns a representation of the dataset. Deep learning models extract important features by iteratively transforming the data, "going deeper" toward meaningful patterns in the dataset with each transformation. However, a large dataset is needed for the utilization of complex healthcare data.

Using transfer learning can potentially solve this problem by bridging the source and target domains vis-a-vis learning invariant feature representations from the source domain. It uses learning from one task on to another task without the requirement of learning from scratch and in this way it directly addresses the smarter parameter initialization point for training the neural network. Using transfer learning, especially in the Healthcare area which suffers from the imbalanced dataset and small labeled dataset improves AUC which has a meaningful interpretation of disease classification from healthy subjects.

In this study, we have trained a model on multiple balanced breast cancer datasets, and transfer the knowledge learned to predict benign/malignant breast cancer status from a severely an imbalanced local dataset with high AUC. We have trained a model on diabetes readmission dataset, and transfer the knowledge learned to predict diabetic/non-diabetic status with higher accuracy. Our transfer learning approach is equally competitive with state-of-the-art methods for varying degrees of imbalanced datasets. We also demonstrated that feature based transfer learning works better compared to network based transfer learning in the case of diabetes prediction. Overall, our results indicate, the the proposed approach resulted in the best classification results compared to other machine algorithms.

These source models are well documented and publicly available in GitHub. We believe that this work provides researchers and professionals with state-of-the-art knowledge on transfer learning with computational intelligence and will provide guidelines about how to develop and apply transfer learning over structured healthcare data to support users in various health-related decisions.

3.13 Acknowledgements

This work is supported by the Mayo Clinic Robert D. and Patricia E. Kern Center for the Science of Health Care Delivery. Dr. McCoy is supported by the National Institute of Diabetes and Digestive and Kidney Diseases of the National Institutes of Health under Award Number K23DK114497. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Table 3.1: Statistical analysis of target MMHS data (Logistics Regression)

Features	Number	Benign(N=14899)	Malignant(n=487)	Total(N=15386)	p-value	coef	OR (30.5)
BI-RADS Breast Density Changes	1						
First Month					0.000	0.017	
1		3652(24.5%)	74(15.2%)	3726 (24.2%)			0.020
2		6175(41.4%)	215(44.1%)	6390(41.5%)			0.034
3		4339(29.1%)	179(36.8%)	4518(29.4%)			0.041
4		733(4.9%)	19(3.9%)	752(4.9%)			0.025
Second Month	2				0.000	0.008	
1		3434 (23.0%)	73(15.0%)	3507(22.8%)			0.021
2		6437(43.2%)	226(46.4%)	6663(43.4%)			0.035
3		4364(29.3%)	166(34.1%)	4530(29.4%)			0.038
4		664(4.5%)	22(4.5%)	686(4.5%)			0.033
Third Month	3				0.83	0.0007	
1		3193(21.4%)	86(17.7%)	3279(21.3%)			0.026
2		6459(43.4%)	200(41.1%)	6659(43.4%)			0.030
3		4633(31.1%)	178(36.6%)	4811(31.3%)			0.038
4		614(4.1%)	23(4.7%)	637(4.1%)			0.037
Fourth Month	4				0.78	-0.0009	
1		2990(20.1%)	89(18.3%)	3079(20.0%)			0.029
2		6638(44.6%)	201(41.3%)	6839(44.4%)			0.030
3		4702(31.6%)	177(36.3%)	4879(31.7%)			0.037
4		569(3.8%)	20(4.1%)	589(3.8%)			0.035
Fifth Month	5				0.98	-0.0007	
1		2809(18.9%)	88(18.1%)	2897(18.8%)			0.031
2		6740(45.2%)	214(43.9%)	6954(45.2%)			0.031
3		4812(32.3%)	163(33.5%)	4975(32.3%)			0.035
4		538(3.6%)	22(4.5%)	560(3.6%)			0.040
Six Month	6				0.13	0.003	
1		2602(17.5%)	89(18.3%)	2691(17.5%)			0.034
2		6858(46.0%)	222(45.6%)	7080(46.0%)			0.032
3		4899(32.9%)	149(30.6%)	5048(32.8%)			0.030
4		540 (3.6%)	27(5.5%)	567(3.7%)			0.05
Family History of BC	7	2698(18.1%)	119(24.4%)	2817(18.3%)	0.007	0.009	
Age at Enrollment	8	56.66(11.31)	58.82(10.30)	56.73(11.29)	0.000	0.037	
Autofluorescence Bronchoscopy Group	9				0.000	0.005	
0		1970(13.2%)	65 (13.3%)	2035(13.2%)			0.032
1		2267(15.2%)	69(14.2%)	2336(15.2%)			0.030
2		5801(38.9%)	193(39.6%)	5994(39.0%)			0.033
3		3308(22.2%)	107(22.0%)	3415(22.2%)			0.032
4		1445(9.7%)	47(9.7%)	1492(9.7%)			0.032
9		108(0.7%)	6(1.2%)	114(0.7%)			0.05
Age at Menarche	10				0.001	0.002	
1		2503(16.8%)	80 (16.4%)	2583(16.8%)			0.031
2		7954 (53.4%)	283(58.1%)	8237(53.5%)			0.035
3		3481 (23.4%)	93 (19.1%)	3574(23.3%)			30.026
8		866(5.8%)	27(5.5%)	893(5.8%)			0.031
9		95 (0.6%)	4(0.8%)	99(0.6%)			0.042
Menopause at Enrollment	11	9081(61.0%)	346(71.0%)	9427(61.3%)	0.20	0.004	
BMI at Enrollment	12	28.14(6.42%)	28.39(6.43%)	28.14(6.42%)	0.000	0.004	
BI-RADS at Enrollment	13				0.000	0.0151	
1		3268(21.9%)	73(15.9%)	3341(21.7%)			0.022
2		5865(39.4%)	194(39.8%)	6059(39.4%)			0.033
3		4731(31.8)	176(36.1%)	4907(31.9%)			0.037
4		1035(6.9%)	44(9.0%)	1079(7.0%)			0.042
Age at Diagnosis	14				0.000	-0.035	
0	30	352(2.3%)	0	352			0
1	40	4253(28.5%)	28(5.74%)	4281(27.8%)			0.006
2	50	4335(29.09%)	110(22.5%)	4445 (28.8%)			0.025
3	60	3492(23.4%)	167 (34.2%)	3659(23.7%)			0.047
4	70	2031(13.63%)	131 (26.8%)	2161 (14%)			0.064
5+	80+	435(2.8%)	51 (9.8%)	485(3%)			0.11

Table 3.2: Statistical analysis of target MMD data (Logistics Regression)

Features	Number	Benign(N=427)	Malignant(n=404)	Total(N=831)	p-value	coef	OR (1.05)
BI-RADS Breast Density Changes	0				0.001	0.023	
0		2(0.4%)	3(0.7%)	5(1.2%)			1.5
2		7 (1.6)	0(0%)	7 (0.8%)			>7
3		20 (4.6%)	4(0.9%)	24(2.8%)			5
4		364 (85)	100(24%)	464(55%)			3.6
5		31 (7)	285 (70%)	316 (38%)			0.1
6		2 (0.4)	7(1.7%)	9(1%)			0.2
Age	1				0.000	0.006	
0		5 (1.17%)	0	5(0.6%)			>5
1		35 (8.1%)	1(0.2%)	36(4.3%)			35
2		64(14.9%)	13 (3.2%)	77(9.2%)			4.9
3		108(25.2%)	51 (12.6%)	159 (19.1%)			2
4		117 (27%)	95 (23%)	212(25.5%)			1.2
5		75 (17.5%)	128 (31.6%)	203 (24.4%)			0.5
6		20 (4.6%))	83 (20.5%)	103 (12.3%)			0.2
7		4 (0.9%)	33 (8.1%)	37 (4.4%)			0.1
8		0	3 (0.7%)	3 (0.3%)			0
Shape	2				0.000	0.094	
1		159(37.5%)	33(8.1%)	192(23%)			4.8
2		149 (34.8 %)	30 (7.4%)	179(21.5%)			4.9
3		39 (9.1%)	42 (10.3%)	81(9.7%)			0.9
4		81 (1.8%)	300 (74%)	381 (45%)			0.2
Margin	3				0.000	0.095	
1		283 (66.2%)	39 (9.6%)	322(38%)			7.2
2		8 (1.9%)	13(3.2%)	21(2.5%)			0.6
3		39 (9.1%)	68 (%)	107 (12.8%)			0.5
4		76 (17.7%)	176 (%)	252 (30%)			0.43
5		20 (4.6%)	106 (%)	126 (15%)			0.18
Density	4				0.000	-0.167	
1		6(1.4%)	5 (1.2%)	11(1.3%)			1.2
2		40 (9.3%)	21(5.1%)	65 (7.8%)			1.9
3		379 (88%)	375 (92%)	754(90%)			1.0
4		4 (0.9%)	4 (0.9%)	8(0.9%)			1

Table 3.3: Parameters of the Breast Cancer Network

l	W^l	$\mathbf{b^l}$
1	$R^{N^0 \times N^1}$ (30×100)	R^{N^1} (100)
2	$R^{N^1 \times N^2}$ (100×50)	R^{N^2} (50)
3	$R^{N^2 \times N^3}$ (50×30)	R^{N^3} (30)
4	$R^{N^3 \times N^4}$ (30×1)	R^{N^4} (1)

Table 3.4: Hyperparameters of the Breast Cancer Network

Activation	Optimizer	Patience	Epoch	Batch	Drop-out	Early-Stop
ELU	Nadam	300	1000	50	0.15	Min-val-loss

Table 3.5: DTL models

a	5L-2T-Freeze	Top 2 Layers transferred from the source model to retrain network on the target model. Transferred layers are frozen during training.
b	5L-2T-UNFreeze	Top 2 Layers transferred from the source model to retrain source network on the target model. Transferred layers are UNfrozen during training.
c	5L-1T-UNFreeze	Top Layer transferred from the source model and 2 hidden layers added to retrain source network on the target model. Transferred layer is UNfrozen during training
d	7L-2T-Freeze	Top 2 Layers transferred from the source model and 3 hidden layers added to retrain source network on the target model. Transferred layers are frozen during training

Table 3.6: Performance of models on hold out fold of MMHS

Classifiers	**AUC-cv**	**Sens-cv**	**F1-cv**
LR	0.59	0.20	0.30
RF	0.61	0.09	0.05
XGBoost	0.68	0.10	0.15
DNN-5L	0.65	0.32	0.37
DNN-7L	0.66	0.32	0.37
LR-sampling	0.65	0.20	0.31
RF-sampling	0.53	0.47	0.07
SMOTEBoost	0.67	0.36	**0.44**
RUSBoost	0.76	0.74	0.17
5L-2T-Freeze	0.80	0.75	0.29
5L-2T-UNFreeze	**0.81**	0.75	0.28
5L-1T-UNFreeze	0.80	0.71	0.27
7L-2T-Freeze	0.76	**0.81**	0.19

Table 3.7: Features of Diabetic Readmission Dataset

Features	SubSets
Inpatient encounter	Gender Age Weight Admission type Race
Diabetic encounter	Payer code Medical specialty Number of diagnosis Admission source Discharge disposition
Length of stay	Number of lab procedures Number of inpatient visits Number of outpatient visits Time in hospital Number of emergency visits
Laboratory tests	Diagnosis 1 Diagnosis 2 Diagnosis 3 Glucose serum test result A1c test result
Medications	Number of procedures Change of medications Diabetes medications 24 features for medications Number of medications

Table 3.8: Parameters of the Diabetes Network using Feature based Transfer learning

l	W^l	$\mathbf{b}^l$
1	$R^{N^0 \times N^1}(10 \times 49)$	R^{N^1} (49)
2	$R^{N^1 \times N^2}(49 \times 30)$	R^{N^2} (30)
3	$R^{N^2 \times N^3}(30 \times 10)$	R^{N^3} (10)
4	$R^{N^3 \times N^4}(10 \times 7)$	R^{N^4} (7)
5	$R^{N^4 \times N^5}(7 \times 1)$	R^{N^5} (1)

Table 3.9: Hyperparameters of the Diabetes network

Activate	Optimizer	Patience	Epoch	Batch	Drop-out	Early-stop
ELU	Adadelta	80	100	50	0.15	Min-val-loss

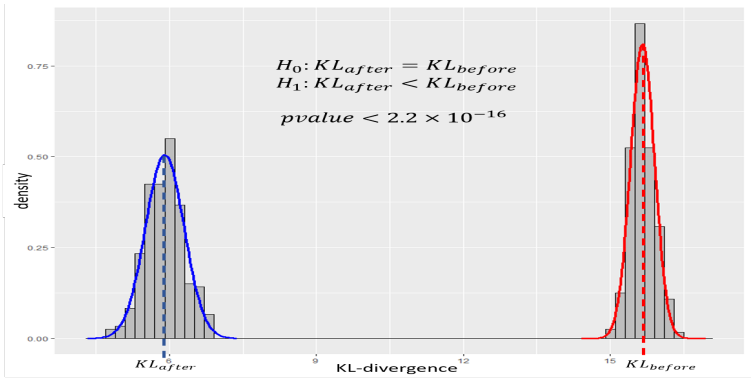

Figure 3.7: Two-sample KL-divergence testing.

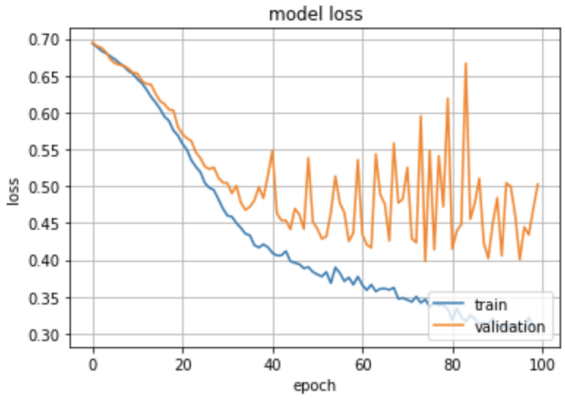

Figure 3.8: BeforeTransfer

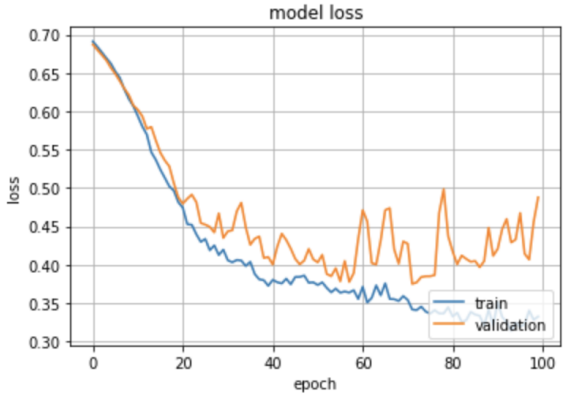

Figure 3.9: AfterTransfer

Table 3.10: Performance of models on diabetes dataset

Classifiers	AUC	Sens	Acc	F1	AUC-cv	Sens-cv	Acc-cv	F1-cv
Logistic-regression	0.72	0.78	0.77	0.77	0.70	0.77	0.75	0.76
LDA-classifier	0.74	0.78	0.78	0.78	0.71	0.78	0.75	0.77
Decision-Tree	0.82	0.82	0.82	0.82	0.65	0.72	0.75	0.71
Random-Forest	0.91	0.93	0.93	0.93	0.67	0.73	0.74	0.73
XGBoost	0.97	0.98	0.97	0.98	0.71	0.75	0.77	0.74
DNN-5L	0.78	0.70	0.82	0.72	0.75	0.68	0.79	0.71
DNN-6L	0.80	0.74	0.83	0.75	0.79	0.72	0.81	0.73
DTL-5L-Feature	0.86	0.79	0.88	0.82	**0.85**	**0.80**	**0.86**	**0.81**
6L-2T-Freeze	0.82	0.77	0.83	0.76	0.81	0.76	0.83	0.75
6L-2T-UNFreeze	0.82	0.79	0.85	0.77	0.82	0.75	0.85	0.75
7L-2T-Freeze	0.80	0.75	0.82	0.75	0.80	0.73	0.82	0.74
7L-2T-UNFreeze	0.82	0.78	0.85	0.77	0.80	0.72	0.82	0.74

Chapter 4

DISCUSSION AND CONCLUSION

EHR systems include structured data (demographic, diagnostic, physical exam, sensor measurements, vital signs, laboratory test) and unstructured data (notes charted by physicians, images, observations and more). Most deep learning uses unstructured data in EHR because obtaining structured data is expensive and time-consuming. Although there are no guidelines about the minimum number of training sets of labeled data, more data can make more accurate machine learning models. Specifically using more data, deep learning has shown promising results for various clinical prediction tasks such as diagnosis, mortality prediction, predicting the duration of stay in the hospital, etc. Healthcare is one of the most common domains suffering from scarcity of labeled data and imbalanced data, and addressing them is an urgent priority. I have done two case studies on structured healthcare data, case study on breast cancer dataset and diabetes dataset. In this dissertation, I introduced my endeavors to address each of the above issues, notably for healthcare data.

In chapter 2, I applied deep transfer learning to solve the problem of the imbalanced dataset and small labeled dataset. This deep transfer learning models use one or multiple source data and transfer knowledge from source data to target data to improve the classification accuracy.

Data scarcity can be overcome by supplementing examples from the target population with a "source" population, for which data are more abundant, in a statistical process known as transfer learning. Transfer learning tackles the problem of leveraging data from a related source task to improve performance on a target task. Auxiliary data tend to have the greatest impact when the number of target training samples is small and there is overlap in shared feature space. Transfer Learning approaches have been applied successfully to medical data. In chapter 3, I have done a meta-analysis to analyze the status of healthcare-related TL studies in terms of the study targets, transfer learning (TL) model(s) used, healthcare data, type of study area, and level of classification accuracy achieved. Subsequently, a detailed review is conducted to describe/discuss how TL has been applied for improving the accuracy of diagnosis in healthcare including images, text, audio, video and structured EHR data classification. In the end, a conclusion regarding the current state-of-the-art methods, a critical conclusion on open challenges, and directions for future research are presented.

Finally, I demonstrated that our transfer learning approach is equally competitive with state-of-the-art methods for varying degrees of imbalanced datasets. Overall, our results indicate, the proposed approach resulted in the best classification results compared to other machine algorithms. Based on this, I conclude that transfer learning techniques can substantially alleviate the burdensome, site-specific data collection requirements for producing effective clinical classifiers. Furthermore, the resulting classifier's performance may be superior to that of the otherwise comparable, non–transfer-trained classifier.

Appendix A

APPENDIX

Table 5.1 presents the mean of the 10 initial features. Coefficients help us to determine the importance of features in the regression model. The p-value for each independent variable tests the null hypothesis that the variable has no correlation with the dependent variable.

A.1 KL Divergence of original data

We compared KL divergence of readmission of diabetic patients and Pima Inidan diabetes before and after encoding. Considering that the number of instances for two dataset were different, we randomly removed some instances from readmission of diabetic patients dataset to match the number with the other dataset. Finally we had 768 instances for both datasets.

We computed KL Divergence for categorical and continuous variables of two datasets separately. (Needs to be noted that some of the features of categorical data are shown here not all)

	Pregnancies	Glucose	BloodPressure	SkinThickness	Insulin	BMI	DiabetesPedigreeFunctior	Age
age	618.728129	359.230543	998.9165904	55.05913846	418.8577614	275.807223	11033.59374	**106.2445**
time_in_hospital	6882.08774	99.9675198	1325.250709	4901.27116	2098.980888	658.975499	23105.63782	203.7901
num_lab_procedures	5094.58703	87.9509179	1135.006757	3870.363017	1801.154649	544.288322	23340.07043	203.7901
num_procedures	10294.7282	115.397761	1560.543974	6205.386799	2481.154085	782.212667	14755.47502	203.7901
num_medications	5188.19229	90.8457518	1186.890595	4115.918242	1867.092942	590.580636	23588.22767	203.7901
number_diagnoses	12816.0108	109.869371	1468.582451	5734.66107	2352.675022	713.124588	**472.4245515**	203.7901
nummed	17927.5875	117.733971	1554.476523	6441.468082	2597.349102	687.595464	11003.07603	203.7901
number_outpatient_log1p	19555.1453	147.145631	2084.754502	8942.377903	3238.848765	1146.27838	10603.18536	203.7901
service_utilization_log1p	16442.5597	140.751416	1974.117402	8358.014566	3079.650002	1059.62558	6934.77598	203.7901
num_medications.time_in_hospital	924.519184	163.26667	373.1326456	248.982947	122.8298529	784.864215	5560.53625	604.2785
num_medications.num_procedures	2006.4034	71.445498	**69.70728065**	389.5516734	530.7675473	436.501703	4509.308341	608.4354
time_in_hospital.num_lab_procedures	1549.92856	186.69964	348.3300811	420.9026353	32.71145735	638.189005	6407.963068	481.2208
num_medications.num_lab_procedures	695.583687	115.251355	162.8483078	231.9611558	78.73064911	316.016111	7331.218402	289.7286
num_medications.number_diagnoses	**504.117403**	156.906746	419.5071857	281.6367066	55.63179743	608.01579	7708.726461	464.3462
age.number_diagnoses	629.107209	84.3004258	123.3174141	119.6349861	**20.3230128**	212.625574	10264.15692	175.8125
change.num_medications	2385.00657	**35.6027242**	598.258114	2057.604314	1142.405035	**14.0939785**	1040.928541	487.8484
number_diagnoses.time_in_hospital	893.758247	62.6184665	240.3885975	**52.02764502**	355.0043981	706.657349	5744.182138	476.9125
num_medications.numchange	5941.78083	66.77899	976.0998608	3397.047113	1557.212467	295.921754	4966.991567	452.0563

Figure A.1: BeforeTransfer

	Glucose.BloodPressure	Age.Glucose
metformin	4291.897998	7193.8296
repaglinide	56887.65616	69332.0585
nateglinide	67433.29624	81612.2143
glipizide	12916.19794	17745.3474
glyburide	15475.34813	20802.4765
pioglitazone	23024.38495	29739.8205
rosiglitazone	25207.15821	32309.4372
acarbose	86725.27663	104062.335
tolazamide	118247.9225	140730.485
insulin	1997.108407	1036.29493
glyburide.metformin	67652.48277	81867.3685
number_inpatient_log1p	20061.00536	26242.3897
gender_1	1006.329367	335.726895
admission_type_id_3	5101.049537	8213.76356
admission_type_id_5	15475.34813	20802.4765
discharge_disposition_id_2	5560.910817	8788.94635
discharge_disposition_id_7	71664.37133	86537.1412
discharge_disposition_id_10	107362.3063	128068.906
discharge_disposition_id_18	30908.08299	39001.6011
admission_source_id_4	25447.46526	32592.0463
admission_source_id_7	2248.316963	1206.98975
admission_source_id_9	24190.09085	31112.7247
max_glu_serum_0	44831.39335	55277.3781
max_glu_serum_1	45887.7786	56509.8568
A1Cresult_0	28416.18231	36079.3367
A1Cresult_1	13012.74876	17861.048
level1_diag1_1.0	**79.36328553**	**1248.27303**

Figure A.2: AfterTransfer

Proposed Deep Transfer Learning for Diabetes using Mapping based transfer learning is presented here.

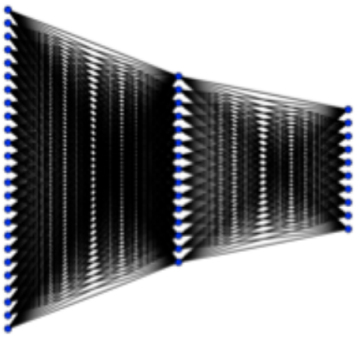

Figure A.3: Source encoder

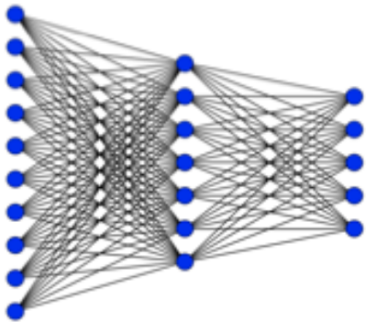

Figure A.4: Target encodere

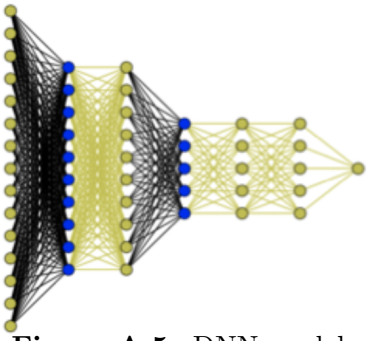

Figure A.5: DNN model

Table A.1: Descriptive statistics of combined WDBC and WPBC source data (Logistics Regression2

Features	Number	Benign(N=508)	Malignant(n=259)	Total(N=767)	p-value	coef
Radius	1				0.02	-2.1
Mean		13.63	17.63	14.97		
SD		3.18	3.24	3.72		
Min		6.99	10.95	6.98		
Max		24.63	28.11	28.11		
Compactness	2				0.78	-0.01
Mean		19.25	21.63	20.06		
SD		4.62	3.75	4.49		
Min		9.71	10.38	9.71		
Max		39.2	39.28	39.28		
Texture	3				0.13	1.68
Mean		88.38	116.49	97.87		
SD		21.77	22.13	25.6		
Min		43.79	71.9	43.7		
Max		166.20	188.50	188.5		
Concavity	4				0.35	0.35
Mean		602.5	998.5	736.2		
SD		301.8	374.9	377.8		
Min		143.5	361.6	143.5		
Max		1841.0	2501	2501		
Perimeter	5				0.6	0.02
Mean		0.095	0.102	0.097		
SD		0.014	0.012	0.013		
Min		0.052	0.073	0.052		
Max		0.16	0.144	0.163		
Concave points	6				0.19	-0.16
Mean		0.098	0.144	0.11		
SD		0.049	0.051	0.054		
Min		0.019	0.046	0.019		
Max		0.311	0.34	0.34		
Area	7				0.91	-0.017
Mean		0.078	0.16	0.106		
SD		0.073	0.072	0.082		
Min		0	0.023	0		
Max		0.42	0.42	0.42		
Symmetry	8				0.66	0.057
Mean		0.043	0.089	0.058		
SD		0.035	0.034	0.041		
Min		0	0.020	0		
Max		0.20	0.20	0.20		
Smoothness	9				0.69	0.014
Mean		0.18	0.19	0.184		
SD		0.027	0.026	0.027		
Min		0.106	0.130	0.106		
Max		0.304	0.304	0.304		
Fractal dimension	10				0.17	-0.09
Mean		0.062	0.062	0.062		
SD		0.006	0.007	0.007		
Min		0.05	0.049	0.049		
Max		0.097	0.097	0.097		

Table A.2: Features of WDBC, WPBC and Combined Source Data

Features of WDBC	Features of WPBC	Features of Combined Dataset
ID number	ID	-
Radius	Radius	Radius
Texture	Texture	Texture
Perimeter	Perimeter	Perimeter
Area	Area	Area
Smoothness	Smoothness	Smoothness
Compactness	Compactness	Compactness
Concavity	Concavity	Concavity
Concave points	Concave points	Concave points
Symmetry	Symmetry	Symmetry
Fractal dimension	Fractal dimension	
-	Time	-
-	Lymph node	-

Table A.3: Tuned Hyper-Parameters of classifiers on diabetes dataset

Classifiers	p1	p2	p3
Logistic-regression	penalty='l2'	solver='lbfgs'	max-iter=500
LDA-classifier	solver='svd'	tol=0.001	component=min(classes-1,feats)
Decision-Tree	min-split=2	max-features=0.7	max-depth=5
Random-Forest	min-split=2	n-estimators=130	max-features=0.7
XGBoost	learn-rate =0.01	n-estimators=1000	max-depth=5
NN-5L	5l	optimizer=adam	activation='elu'
NN-6L	6l	optimizer=adam	activation='elu'
6L-2T-Freeze	6l	optimizer=adam	activation='elu'
6L-2T-UNFreeze	6l	optimizer=adam	activation='elu'
7L-2T-Freeze	7l	optimizer=adam	activation='elu'
7L-2T-UNFreeze	7l	optimizer=adam	activation='elu'

Table A.4: Grid search on Neural network hyperparameters

Learning-rate	optimization	activation	epochs
0.1	Adam	Sigmoid	30
0.01	Nadam	Tanh	50
0.001	Adadelta	Elu	100
0.0001	SGD	ReLU	150
0.00001	RMSprop	Swish	300

REFERENCES

[1] J Henry, Yuriy Pylypchuk, Talisha Searcy, and Vaishali Patel. Adoption of electronic health record systems among us non-federal acute care hospitals: 2008–2015. *ONC data brief*, 35:1–9, 2016.

[2] Ruth A Bush, Cynthia Kuelbs, Julie Ryu, Wen Jiang, and George Chiang. Structured data entry in the electronic medical record: perspectives of pediatric specialty physicians and surgeons. *Journal of medical systems*, 41(5):75, 2017.

[3] Anobel Y Odisho, Mark Bridge, Mitchell Webb, Niloufar Ameli, Renu S Eapen, Frank Stauf, Janet E Cowan, Samuel L Washington III, Annika Herlemann, Peter R Carroll, et al. Automating the capture of structured pathology data for prostate cancer clinical care and research. *JCO clinical cancer informatics*, 3:1–8, 2019.

[4] Jisoo Ahn and Lee Ann Kahlor. No regrets when it comes to your health: Anticipated regret, subjective norms, information insufficiency and intent to seek health information from multiple sources. *Health communication*, pages 1–8, 2019.

[5] Siriwon Taewijit, Thanaruk Theeramunkong, and Mitsuru Ikeda. Distant supervision with transductive learning for adverse drug reaction identification from electronic medical records. *Journal of healthcare engineering*, 2017, 2017.

[6] Nicolas Garcelon, Anita Burgun, Rémi Salomon, and Antoine Neuraz. Electronic health records for the diagnosis of rare diseases. *Kidney International*, 2020.

[7] Rohini R Rao and Krishnamoorthi Makkithaya. Learning from a class imbalanced public health dataset: A cost-based comparison of classifier performance. *International Journal of Electrical and Computer Engineering*, 7(4):2215, 2017.

[8] Gloria Hyun-Jung Kwak and Pan Hui. Deephealth: Deep learning for health informatics. *arXiv preprint arXiv:1909.00384*, 2019.

[9] Jinmeng Jia, Ruiyuan Wang, Zhongxin An, Yongli Guo, Xi Ni, and Tieliu Shi. Rdad: A machine learning system to support phenotype-based rare disease diagnosis. *Frontiers in genetics*, 9:587, 2018.

[10] Toyofumi Fujiwara, Yasunori Yamamoto, Jin-Dong Kim, Orion Buske, and Toshihisa Takagi. Pubcasefinder: A case-report-based, phenotype-driven differential-diagnosis system for rare diseases. *The American Journal of Human Genetics*, 103(3):389–399, 2018.

[11] Rachel L Richesson, Jimeng Sun, Jyotishman Pathak, Abel N Kho, and Joshua C Denny. Clinical phenotyping in selected national networks: demonstrating the need for high-throughput, portable, and computational methods. *Artificial intelligence in medicine*, 71:57–61, 2016.

[12] Brett Kreigh Beaulieu-Jones. Machine learning methods to identify hidden phenotypes in the electronic health record. 2017.

[13] Hamed Hassanzadeh, Mahnoosh Kholghi, Anthony Nguyen, and Kevin Chu. Clinical document classification using labeled and unlabeled data across hospitals. In *AMIA Annual Symposium Proceedings*, volume 2018, page 545. American Medical Informatics Association, 2018.

[14] Jina Ko, Steven N Baldassano, Po-Ling Loh, Konrad Kording, Brian Litt, and David Issadore. Machine learning to detect signatures of disease in liquid biopsies–a user's guide. *Lab on a Chip*, 18(3):395–405, 2018.

[15] Vinayak Athavale, Shreenivas Bharadwaj, Monik Pamecha, Ameya Prabhu, and Manish Shrivastava. Towards deep learning in hindi ner: An approach to tackle the labelled data scarcity. *arXiv preprint arXiv:1610.09756*, 2016.

[16] Mehdi Gheisari, Guojun Wang, and Md Zakirul Alam Bhuiyan. A survey on deep learning in big data. In *2017 IEEE international conference on computational science and engineering (CSE) and IEEE international conference on embedded and ubiquitous computing (EUC)*, volume 2, pages 173–180. IEEE, 2017.

[17] Francois Chollet. The limitations of deep learning. *Deep Learning With Python*, 2017.

[18] Gary Marcus. Deep learning: A critical appraisal. *arXiv preprint arXiv:1801.00631*, 2018.

[19] SJ Pan and Q Yang. A survey on transfer learning. ieee transaction on knowledge discovery and data engineering, 22 (10), 2010.

[20] Chuanqi Tan, Fuchun Sun, Tao Kong, Wenchang Zhang, Chao Yang, and Chunfang Liu. A survey on deep transfer learning. In *International Conference on Artificial Neural Networks*, pages 270–279. Springer, 2018.

[21] Benjamin Shickel, Patrick James Tighe, Azra Bihorac, and Parisa Rashidi. Deep ehr: a survey of recent advances in deep learning techniques for electronic health record (ehr) analysis. *IEEE journal of biomedical and health informatics*, 22(5):1589–1604, 2017.

[22] Feifan Liu, Jinying Chen, Abhyuday Jagannatha, and Hong Yu. Learning for biomedical information extraction: Methodological review of recent advances. *arXiv preprint arXiv:1606.07993*, 2016.

[23] Cao Xiao, Edward Choi, and Jimeng Sun. Opportunities and challenges in developing deep learning models using electronic health records data: a systematic review. *Journal of the American Medical Informatics Association*, 25(10):1419–1428, 2018.

[24] Trevor Darrell, Marius Kloft, Massimiliano Pontil, Gunnar Rätsch, and Erik Rodner. Machine learning with interdependent and non-identically distributed data (dagstuhl seminar 15152). In *Dagstuhl Reports*, volume 5. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2015.

[25] Elnaz Soleimani and Ehsan Nazerfard. Cross-subject transfer learning in human activity recognition systems using generative adversarial networks. *arXiv preprint arXiv:1903.12489*, 2019.

[26] Geert Litjens, Thijs Kooi, Babak Ehteshami Bejnordi, Arnaud Arindra Adiyoso Setio, Francesco Ciompi, Mohsen Ghafoorian, Jeroen Awm Van Der Laak, Bram Van Ginneken, and Clara I Sánchez. A survey on deep learning in medical image analysis. *Medical image analysis*, 42:60–88, 2017.

[27] Shaveta Dargan, Munish Kumar, Maruthi Rohit Ayyagari, and Gulshan Kumar. A survey of deep learning and its applications: A new paradigm to machine learning. *Archives of Computational Methods in Engineering*, pages 1–22, 2019.

[28] Sinno Jialin Pan and Qiang Yang. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2009.

[29] Jenna Wiens, John Guttag, and Eric Horvitz. A study in transfer learning: leveraging

data from multiple hospitals to enhance hospital-specific predictions. *Journal of the American Medical Informatics Association*, 21(4):699–706, 2014.

[30] Zhenghui Wang, Yanru Qu, Liheng Chen, Jian Shen, Weinan Zhang, Shaodian Zhang, Yimei Gao, Gen Gu, Ken Chen, and Yong Yu. Label-aware double transfer learning for cross-specialty medical named entity recognition. *arXiv preprint arXiv:1804.09021*, 2018.

[31] Lisa Torrey and Jude Shavlik. Transfer learning. In *Handbook of research on machine learning applications and trends: algorithms, methods, and techniques*, pages 242–264. IGI global, 2010.

[32] Hong-Wei Ng, Viet Dung Nguyen, Vassilios Vonikakis, and Stefan Winkler. Deep learning for emotion recognition on small datasets using transfer learning. In *Proceedings of the 2015 ACM on international conference on multimodal interaction*, pages 443–449, 2015.

[33] Akram Farhadi, David Chen, Rozalina McCoy, Christopher Scott, John A Miller, Celine M Vachon, and Che Ngufor. Breast cancer classification using deep transfer learning on structured healthcare data. In *2019 IEEE International Conference on Data Science and Advanced Analytics (DSAA)*, pages 277–286. IEEE, 2019.

[34] Zhaocheng Wang, Lan Du, Jiashun Mao, Bin Liu, and Dongwen Yang. Sar target detection based on ssd with data augmentation and transfer learning. *IEEE Geoscience and Remote Sensing Letters*, 16(1):150–154, 2018.

[35] Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. Heterogeneous transfer learning for image classification. In *Twenty-Fifth AAAI Conference on Artificial Intelligence*, 2011.

[36] Yuntao Du, Zhiwen Tan, Qian Chen, Yi Zhang, and Chongjun Wang. Homogeneous online transfer learning with online distribution discrepancy minimization. *arXiv preprint arXiv:1912.13226*, 2019.

[37] Stefanie Utech, Radivoje Prodanovic, Angelo S Mao, Raluca Ostafe, David J Mooney, and David A Weitz. Microfluidic generation of monodisperse, structurally homogeneous alginate microgels for cell encapsulation and 3d cell culture. *Advanced healthcare materials*, 4(11):1628–1633, 2015.

[38] Khalid M Hosny, Mohamed A Kassem, and Mohamed M Foaud. Skin cancer classification using deep learning and transfer learning. In *2018 9th Cairo International Biomedical Engineering Conference (CIBEC)*, pages 90–93. IEEE, 2018.

[39] Afonso Menegola, Michel Fornaciali, Ramon Pires, Sandra Avila, and Eduardo Valle. Towards automated melanoma screening: Exploring transfer learning schemes. *arXiv preprint arXiv:1609.01228*, 2016.

[40] Zabir Al Nazi and Tasnim Azad Abir. Automatic skin lesion segmentation and melanoma detection: Transfer learning approach with u-net and dcnn-svm. In *Proceedings of International Joint Conference on Computational Intelligence*, pages 371–381. Springer, 2020.

[41] Rajesh Mehra et al. Breast cancer histology images classification: Training from scratch or transfer learning? *ICT Express*, 4(4):247–254, 2018.

[42] Richa Agarwal, Oliver Diaz, Xavier Lladó, Moi Hoon Yap, and Robert Martí. Automatic mass detection in mammograms using deep convolutional neural networks. *Journal of Medical Imaging*, 6(3):031409, 2019.

[43] MS Basanth and Rajashree Shettar. Transfer learning on pre-trained deep convolu-

tional neural network for classification of masses in mammograms. *IOSR J Comput Eng*, 19(50):e5, 2017.

[44] Sara Hosseinzadeh Kassani, Peyman Hosseinzadeh Kassani, Michal J Wesolowski, Kevin A Schneider, and Ralph Deters. Breast cancer diagnosis with transfer learning and global pooling. *arXiv preprint arXiv:1909.11839*, 2019.

[45] Ravi K Samala, Heang-Ping Chan, Lubomir M Hadjiiski, Mark A Helvie, Kenny H Cha, and Caleb D Richter. Multi-task transfer learning deep convolutional neural network: application to computer-aided diagnosis of breast cancer on mammograms. *Physics in Medicine & Biology*, 62(23):8894, 2017.

[46] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Caleb D Richter, and Kenny H Cha. Breast cancer diagnosis in digital breast tomosynthesis: effects of training sample size on multi-stage transfer learning using deep neural nets. *IEEE Transactions on Medical Imaging*, 38(3):686–696, 2018.

[47] Fan Jiang, Hui Liu, Shaode Yu, and Yaoqin Xie. Breast mass lesion classification in mammograms by transfer learning. In *Proceedings of the 5th international conference on bioinformatics and computational biology*, pages 59–62, 2017.

[48] Hiba Chougrad, Hamid Zouaki, and Omar Alheyane. Convolutional neural networks for breast cancer screening: transfer learning with exponential decay. *arXiv preprint arXiv:1711.10752*, 2017.

[49] Michal Byra, Tomasz Sznajder, Danijel Korzinek, Hanna Piotrzkowska-Wróblewska, Katarzyna Dobruch-Sobczak, Andrzej Nowicki, and Krzysztof Marasek. Impact of ultrasound image reconstruction method on breast lesion classification with neural transfer learning. *arXiv preprint arXiv:1804.02119*, 2018.

[50] Sulaiman Vesal, Nishant Ravikumar, AmirAbbas Davari, Stephan Ellmann, and Andreas Maier. Classification of breast cancer histology images using transfer learning. In *International conference image analysis and recognition*, pages 812–819. Springer, 2018.

[51] Hao Pang, Wenjie Lin, Cong Wang, and Chen Zhao. Using transfer learning to detect breast cancer without network training. In *2018 5th IEEE International Conference on Cloud Computing and Intelligence Systems (CCIS)*, pages 381–385. IEEE, 2018.

[52] Quan Chen, Shiliang Hu, Peiran Long, Fang Lu, Yujie Shi, and Yunpeng Li. A transfer learning approach for malignant prostate lesion detection on multiparametric mri. *Technology in cancer research & treatment*, 18:1533033819858363, 2019.

[53] Yixuan Yuan, Wenjian Qin, Mark Buyyounouski, Bulat Ibragimov, Steve Hancock, Bin Han, and Lei Xing. Prostate cancer classification with multiparametric mri transfer learning model. *Medical physics*, 46(2):756–765, 2019.

[54] Tiantian Fang. A novel computer-aided lung cancer detection method based on transfer learning from googlenet and median intensity projections. In *2018 IEEE International Conference on Computer and Communication Engineering Technology (CCET)*, pages 286–290. IEEE, 2018.

[55] Xin Zhen, Jiawei Chen, Zichun Zhong, Brian Hrycushko, Linghong Zhou, Steve Jiang, Kevin Albuquerque, and Xuejun Gu. Deep convolutional neural network with transfer learning for rectum toxicity prediction in cervical cancer radiotherapy: a feasibility study. *Physics in Medicine & Biology*, 62(21):8246, 2017.

[56] Nils Gessert, Marcel Bengs, Lukas Wittig, Daniel Drömann, Tobias Keck, Alexander Schlaefer, and David B Ellebrecht. Deep transfer learning methods for colon cancer

classification in confocal laser microscopy images. *International journal of computer assisted radiology and surgery*, 14(11):1837–1845, 2019.

[57] Eduardo Ribeiro, Andreas Uhl, Georg Wimmer, and Michael Häfner. Exploring deep learning and transfer learning for colonic polyp classification. *Computational and mathematical methods in medicine*, 2016, 2016.

[58] Sjors Van Riel, Fons Van Der Sommen, Sveta Zinger, Erik J Schoon, and Peter HN de With. Automatic detection of early esophageal cancer with cnns using transfer learning. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1383–1387. IEEE, 2018.

[59] Jin Li, Peng Wang, Yanzhao Li, Yang Zhou, Xiaolong Liu, and Kuan Luan. Transfer learning of pre-trained inception-v3 model for colorectal cancer lymph node metastasis classification. In *2018 IEEE International Conference on Mechatronics and Automation (ICMA)*, pages 1650–1654. IEEE, 2018.

[60] Xiaoqi Liu, Chengliang Wang, Yao Hu, Zhuo Zeng, Jianying Bai, and Guobin Liao. Transfer learning with convolutional neural network for early gastric cancer classification on magnifiying narrow-band imaging images. In *2018 25th IEEE International Conference on Image Processing (ICIP)*, pages 1388–1392. IEEE, 2018.

[61] Muazzam Maqsood, Faria Nazir, Umair Khan, Farhan Aadil, Habibullah Jamal, Irfan Mehmood, and Oh-young Song. Transfer learning assisted classification and detection of alzheimer's disease stages using 3d mri scans. *Sensors*, 19(11):2645, 2019.

[62] Rachna Jain, Nikita Jain, Akshay Aggarwal, and D Jude Hemanth. Convolutional neural network based alzheimer's disease classification from magnetic resonance brain images. *Cognitive Systems Research*, 57:147–159, 2019.

[63] Ke Zhou, Wenguang He, Yonghui Xu, Gangqiang Xiong, and Jie Cai. Feature selection and transfer learning for alzheimer's disease clinical diagnosis. *Applied Sciences*, 8(8):1372, 2018.

[64] Kanghan Oh, Young-Chul Chung, Ko Woon Kim, Woo-Sung Kim, and Il-Seok Oh. Classification and visualization of alzheimer's disease using volumetric convolutional neural network and transfer learning. *Scientific Reports*, 9(1):1–16, 2019.

[65] Marcia Hon and Naimul Mefraz Khan. Towards alzheimer's disease classification through transfer learning. In *2017 IEEE International conference on bioinformatics and biomedicine (BIBM)*, pages 1166–1169. IEEE, 2017.

[66] Naimul Mefraz Khan, Nabila Abraham, and Marcia Hon. Transfer learning with intelligent training data selection for prediction of alzheimer's disease. *IEEE Access*, 7:72726–72735, 2019.

[67] Bo Cheng, Mingxia Liu, Daoqiang Zhang, Brent C Munsell, and Dinggang Shen. Domain transfer learning for mci conversion prediction. *IEEE Transactions on Biomedical Engineering*, 62(7):1805–1817, 2015.

[68] Bo Cheng, Mingxia Liu, Dinggang Shen, Zuoyong Li, Daoqiang Zhang, Alzheimer's Disease Neuroimaging Initiative, et al. Multi-domain transfer learning for early diagnosis of alzheimer's disease. *Neuroinformatics*, 15(2):115–132, 2017.

[69] Genevieve CY Chan, Awais Muhammad, Syed AA Shah, Tong B Tang, Cheng-Kai Lu, and Fabrice Meriaudeau. Transfer learning for diabetic macular edema (dme) detection on optical coherence tomography (oct) images. In *2017 IEEE International Conference on Signal and Image Processing Applications (ICSIPA)*, pages 493–496. IEEE, 2017.

[70] Carson Lam, Darvin Yi, Margaret Guo, and Tony Lindsey. Automated detection of

diabetic retinopathy using deep learning. *AMIA Summits on Translational Science Proceedings*, 2018:147, 2018.

[71] Sri Phani Krishna Karri, Debjani Chakraborty, and Jyotirmoy Chatterjee. Transfer learning based classification of optical coherence tomography images with diabetic macular edema and dry age-related macular degeneration. *Biomedical optics express*, 8(2):579–592, 2017.

[72] Xiaogang Li, Tiantian Pang, Biao Xiong, Weixiang Liu, Ping Liang, and Tianfu Wang. Convolutional neural networks based transfer learning for diabetic retinopathy fundus image classification. In *2017 10th International Congress on Image and Signal Processing, BioMedical Engineering and Informatics (CISP-BMEI)*, pages 1–11. IEEE, 2017.

[73] Kh Tohidul Islam, Sudanthi Wijewickrema, and Stephen O'Leary. Identifying diabetic retinopathy from oct images using deep transfer learning with artificial neural networks. In *2019 IEEE 32nd International Symposium on Computer-Based Medical Systems (CBMS)*, pages 281–286. IEEE, 2019.

[74] Michelle Yuen Ting Yip, Zhan Wei Lim, Gilbert Lim, Nguyen Duc Quang, Haslina Hamzah, Jinyi Ho, Valentina Bellemo, Yuchen Xie, Xin Qi Lee, Mong Li Lee, et al. Enhanced detection of referable diabetic retinopathy via dcnns and transfer learning. In *Asian Conference on Computer Vision*, pages 282–288. Springer, 2018.

[75] Daniel H Kim, Huub Wit, and Mark Thurston. Artificial intelligence in the diagnosis of parkinson's disease from ioflupane-123 single-photon emission computed tomography dopamine transporter scans using transfer learning. *Nuclear medicine communications*, 39(10):887–893, 2018.

[76] Amina Naseer, Monail Rani, Saeeda Naz, Muhammad Imran Razzak, Muhammad Imran, and Guandong Xu. Refining parkinson's neurological disorder identification through deep transfer learning. *Neural Computing and Applications*, 32(3):839–854, 2020.

[77] Angel Cruz-Roa, John Arévalo, Alexander Judkins, Anant Madabhushi, and Fabio González. A method for medulloblastoma tumor differentiation based on convolutional neural networks and transfer learning. In *11th International Symposium on Medical Information Processing and Analysis*, volume 9681, page 968103. International Society for Optics and Photonics, 2015.

[78] Arshia Rehman, Saeeda Naz, Muhammad Imran Razzak, Faiza Akram, and Muhammad Imran. A deep learning-based framework for automatic brain tumors classification using transfer learning. *Circuits, Systems, and Signal Processing*, 39(2):757–775, 2020.

[79] Muhammed Talo, Ulas Baran Baloglu, Özal Yıldırım, and U Rajendra Acharya. Application of deep transfer learning for automated brain abnormality classification using mr images. *Cognitive Systems Research*, 54:176–188, 2019.

[80] Mohsen Ghafoorian, Alireza Mehrtash, Tina Kapur, Nico Karssemeijer, Elena Marchiori, Mehran Pesteie, Charles RG Guttmann, Frank-Erik de Leeuw, Clare M Tempany, Bram van Ginneken, et al. Transfer learning for domain adaptation in mri: Application in brain lesion segmentation. In *International conference on medical image computing and computer-assisted intervention*, pages 516–524. Springer, 2017.

[81] S Deepak and PM Ameer. Brain tumor classification using deep cnn features via transfer learning. *Computers in biology and medicine*, 111:103345, 2019.

[82] Awwal Muhammad Dawud, Kamil Yurtkan, and Huseyin Oztoprak. Application of

deep learning in neuroradiology: Brain haemorrhage classification using transfer learning. *Computational Intelligence and Neuroscience*, 2019, 2019.

[83] Gaobo Liang and Lixin Zheng. A transfer learning method with deep residual network for pediatric pneumonia diagnosis. *Computer methods and programs in biomedicine*, page 104964, 2019.

[84] Cheng Wang, Delei Chen, Lin Hao, Xuebo Liu, Yu Zeng, Jianwei Chen, and Guokai Zhang. Pulmonary image classification based on inception-v3 transfer learning model. *IEEE Access*, 7:146533–146541, 2019.

[85] Qian Wang, Hong Wang, Lutong Wang, and Fengping Yu. Diagnosis of chronic obstructive pulmonary disease based on transfer learning. *IEEE Access*, 8:47370–47383, 2020.

[86] Hoo-Chang Shin, Holger R Roth, Mingchen Gao, Le Lu, Ziyue Xu, Isabella Nogues, Jianhua Yao, Daniel Mollura, and Ronald M Summers. Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning. *IEEE transactions on medical imaging*, 35(5):1285–1298, 2016.

[87] Shingo Mabu, Ami Atsumo, Shoji Kido, Takashi Kuremoto, and Yasushi Hirano. Investigating the effects of transfer learning on roi-based classification of chest ct images: A case study on diffuse lung diseases. *Journal of Signal Processing Systems*, pages 1–7, 2020.

[88] Veronika Cheplygina, Isabel Pino Pena, Jesper Holst Pedersen, David A Lynch, Lauge Sørensen, and Marleen de Bruijne. Transfer learning for multicenter classification of chronic obstructive pulmonary disease. *IEEE journal of biomedical and health informatics*, 22(5):1486–1496, 2017.

[89] Stergios Christodoulidis, Marios Anthimopoulos, Lukas Ebner, Andreas Christe, and Stavroula Mougiakakou. Multisource transfer learning with convolutional neural networks for lung pattern analysis. *IEEE journal of biomedical and health informatics*, 21(1):76–84, 2016.

[90] Ahmed Alghamdi, Mohamed Hammad, Hassan Ugail, Asmaa Abdel-Raheem, Khan Muhammad, Hany S Khalifa, and Ahmed A Abd El-Latif. Detection of myocardial infarction based on novel deep transfer learning methods for urban healthcare in smart cities. *Multimedia Tools and Applications*, pages 1–22, 2020.

[91] Girmaw Abebe Tadesse, Tingting Zhu, Yong Liu, Yingling Zhou, Jiyan Chen, Maoyi Tian, and David Clifton. Cardiovascular disease diagnosis using cross-domain transfer learning. In *2019 41st Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 4262–4265. IEEE, 2019.

[92] Nils Gessert, Markus Heyder, Sarah Latus, Matthias Lutz, and Alexander Schlaefer. Plaque classification in coronary arteries from ivoct images using convolutional neural networks and transfer learning. *arXiv preprint arXiv:1804.03904*, 2018.

[93] Ting Xiao, Lei Liu, Kai Li, Wenjian Qin, Shaode Yu, and Zhicheng Li. Comparison of transferred deep neural networks in ultrasonic breast masses discrimination. *BioMed research international*, 2018, 2018.

[94] Manal Al Ghamdi, Mingqi Li, Mohamed Abdel-Mottaleb, and Mohamed Abou Shousha. Semi-supervised transfer learning for convolutional neural networks for glaucoma detection. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 3812–3816. IEEE, 2019.

[95] Ryo Asaoka, Hiroshi Murata, Kazunori Hirasawa, Yuri Fujino, Masato Matsuura, Atsuya Miki, Takashi Kanamoto, Yoko Ikeda, Kazuhiko Mori, Aiko Iwase, et al. Using

deep learning and transfer learning to accurately diagnose early-onset glaucoma from macular optical coherence tomography images. *American journal of ophthalmology*, 198:136–145, 2019.

[96] Sarfaraz Masood, Tarun Luthra, Himanshu Sundriyal, and Mumtaz Ahmed. Identification of diabetic retinopathy in eye images using transfer learning. In *2017 International Conference on Computing, Communication and Automation (ICCCA)*, pages 1183–1187. IEEE, 2017.

[97] Mark Christopher, Akram Belghith, Christopher Bowd, James A Proudfoot, Michael H Goldbaum, Robert N Weinreb, Christopher A Girkin, Jeffrey M Liebmann, and Linda M Zangwill. Performance of deep learning architectures and transfer learning for detecting glaucomatous optic neuropathy in fundus photographs. *Scientific reports*, 8(1):1–13, 2018.

[98] Juan J Gómez-Valverde, Alfonso Antón, Gianluca Fatti, Bart Liefers, Alejandra Herranz, Andrés Santos, Clara I Sánchez, and María J Ledesma-Carbayo. Automatic glaucoma classification using color fundus images based on convolutional neural networks and transfer learning. *Biomedical optics express*, 10(2):892–913, 2019.

[99] Yiqiang Chen, Xin Qin, Jindong Wang, Chaohui Yu, and Wen Gao. Fedhealth: A federated transfer learning framework for wearable healthcare. *IEEE Intelligent Systems*, 2020.

[100] Dan Meng, Libo Zhang, Guitao Cao, Wenming Cao, Guixu Zhang, and Bing Hu. Liver fibrosis classification based on transfer learning and fcnet for ultrasound images. *Ieee Access*, 5:5804–5810, 2017.

[101] Renjie Ding, Xue Li, Lanshun Nie, Jiazhen Li, Xiandong Si, Dianhui Chu, Guozhong

Liu, and Dechen Zhan. Empirical study and improvement on deep transfer learning for human activity recognition. *Sensors*, 19(1):57, 2019.

[102] Michał Byra, Grzegorz Styczynski, Cezary Szmigielski, Piotr Kalinowski, Łukasz Michałowski, Rafał Paluszkiewicz, Bogna Ziarkiewicz-Wróblewska, Krzysztof Zieniewicz, Piotr Sobieraj, and Andrzej Nowicki. Transfer learning with deep convolutional neural network for liver steatosis assessment in ultrasound images. *International journal of computer assisted radiology and surgery*, 13(12):1895–1903, 2018.

[103] Benjamin Shickel, Martin Heesacker, Sherry Benton, and Parisa Rashidi. Hashtag healthcare: from tweets to mental health journals using deep transfer learning. *arXiv preprint arXiv:1708.01372*, 2017.

[104] Yonghui Xu, Huaqing Min, Qingyao Wu, Hengjie Song, and Bicui Ye. Multi-instance metric transfer learning for genome-wide protein function prediction. *Scientific reports*, 7:41831, 2017.

[105] Edmond Zhang, Quentin Thurier, and Luke Boyle. Improving clinical named-entity recognition with transfer learning. *Studies in health technology and informatics*, 252:182–187, 2018.

[106] K Shaga Devan, Paul Walther, Jens von Einem, Timo Ropinski, Hans A Kestler, and Clarissa Read. Detection of herpesvirus capsids in transmission electron microscopy images using transfer learning. *Histochemistry and cell biology*, 151(2):101–114, 2019.

[107] Xiuli Li, Hao Zhang, Xiaolu Zhang, Hao Liu, and Guotong Xie. Exploring transfer learning for gastrointestinal bleeding detection on small-size imbalanced endoscopy images. In *2017 39th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 1994–1997. IEEE, 2017.

[108] Eric Eaton et al. Set-based boosting for instance-level transfer. In *2009 IEEE International Conference on Data Mining Workshops*, pages 422–428. IEEE, 2009.

[109] YiNan Zhang and MingQiang An. Deep learning-and transfer learning-based super resolution reconstruction from single medical image. *Journal of healthcare engineering*, 2017, 2017.

[110] Xinbo Lv, Yi Guan, and Benyang Deng. Transfer learning based clinical concept extraction on data from multiple sources. *Journal of biomedical informatics*, 52:55–64, 2014.

[111] Azin Asgarian, Parinaz Sobhani, Ji Chao Zhang, Madalin Mihailescu, Ariel Sibilia, Ahmed Bilal Ashraf, and Babak Taati. A hybrid instance-based transfer learning method. *arXiv preprint arXiv:1812.01063*, 2018.

[112] Wenyuan Dai, Qiang Yang, Gui-Rong Xue, and Yong Yu. Boosting for transfer learning. In *Proceedings of the 24th international conference on Machine learning*, pages 193–200, 2007.

[113] Netzahualcoyotl Hernandez, Muhammad Asif Razzaq, Chris Nugent, Ian McChesney, and Shuai Zhang. Transfer learning and data fusion approach to recognize activities of daily life. In *Proceedings of the 12th EAI International Conference on Pervasive Computing Technologies for Healthcare*, pages 227–231, 2018.

[114] Hyunsoo Yoon. *New Statistical Transfer Learning Models for Health Care Applications*. Arizona State University, 2018.

[115] Peter L Bartlett and Mikhail Traskin. Adaboost is consistent. *Journal of Machine Learning Research*, 8(Oct):2347–2368, 2007.

[116] Alexis Bellot and Mihaela Schaar. Boosting transfer learning with survival data from heterogeneous domains. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 57–65, 2019.

[117] Md Abdullah Al Hafiz Khan and Nirmalya Roy. Transact: Transfer learning enabled activity recognition. In *2017 IEEE International Conference on Pervasive Computing and Communications Workshops (PerCom Workshops)*, pages 545–550. IEEE, 2017.

[118] Yuan Shi, Zhenzhong Lan, Wei Liu, and Wei Bi. Extending semi-supervised learning methods for inductive transfer learning. In *2009 Ninth IEEE international conference on data mining*, pages 483–492. IEEE, 2009.

[119] Veronika Cheplygina, Marleen de Bruijne, and Josien PW Pluim. Not-so-supervised: a survey of semi-supervised, multi-instance, and transfer learning in medical image analysis. *Medical image analysis*, 54:280–296, 2019.

[120] Qian Xu and Qiang Yang. A survey of transfer and multitask learning in bioinformatics. *Journal of Computing Science and Engineering*, 5(3):257–268, 2011.

[121] Yi Peng and Li Chen. Multi-feature based transfer function design for 3d medical image visualization. In *2010 3rd International Conference on Biomedical Engineering and Informatics*, volume 1, pages 410–413. IEEE, 2010.

[122] Gyemin Lee, Ilan Rubinfeld, and Zeeshan Syed. Adapting surgical models to individual hospitals using transfer learning. In *2012 IEEE 12th international conference on data mining workshops*, pages 57–63. IEEE, 2012.

[123] Tianjiao Liu, Shuaining Xie, Jing Yu, Lijuan Niu, and Weidong Sun. Classification of thyroid nodules in ultrasound images using deep model based transfer learning and hybrid features. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 919–923. IEEE, 2017.

[124] Rahul Paul, Samuel H Hawkins, Yoganand Balagurunathan, Matthew B Schabath, Robert J Gillies, Lawrence O Hall, and Dmitry B Goldgof. Deep feature transfer learning in combination with traditional features predicts survival among patients with lung adenocarcinoma. *Tomography*, 2(4):388, 2016.

[125] Yizhang Jiang, Dongrui Wu, Zhaohong Deng, Pengjiang Qian, Jun Wang, Guanjin Wang, Fu-Lai Chung, Kup-Sze Choi, and Shitong Wang. Seizure classification from eeg signals using transfer learning, semi-supervised learning and tsk fuzzy system. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 25(12):2270–2284, 2017.

[126] R Mohanasundaram, Ankit Sandeep Malhotra, R Arun, and PS Periasamy. Deep learning and semi-supervised and transfer learning algorithms for medical imaging. In *Deep Learning and Parallel Computing Environment for Bioengineering Systems*, pages 139–151. Elsevier, 2019.

[127] Alban Maxhuni, Pablo Hernandez-Leal, L Enrique Sucar, Venet Osmani, Eduardo F Morales, and Oscar Mayora. Stress modelling and prediction in presence of scarce data. *Journal of biomedical informatics*, 63:344–356, 2016.

[128] J Tahmoresnezhad and S Hashemi. Transductive transfer learning via maximum margin criterion. *Scientia Iranica. Transaction D, Computer Science & Engineering, Electrical*, 23(3):1239, 2016.

[129] Karsten M Borgwardt, Arthur Gretton, Malte J Rasch, Hans-Peter Kriegel, Bernhard Schölkopf, and Alex J Smola. Integrating structured biological data by kernel maximum mean discrepancy. *Bioinformatics*, 22(14):e49–e57, 2006.

[130] Xiaodong Jia, Ming Zhao, Yuan Di, Qibo Yang, and Jay Lee. Assessment of data

suitability for machine prognosis using maximum mean discrepancy. *IEEE transactions on industrial electronics*, 65(7):5872–5881, 2017.

[131] Bo Cheng, Mingxia Liu, Heung-Il Suk, Dinggang Shen, Daoqiang Zhang, Alzheimer's Disease Neuroimaging Initiative, et al. Multimodal manifold-regularized transfer learning for mci conversion prediction. *Brain imaging and behavior*, 9(4):913–926, 2015.

[132] Indra B Nigam. Device for detecting tachycardia using a counter, May 9 2000. US Patent 6,061,592.

[133] Ronald Kemker and Christopher Kanan. Self-taught feature learning for hyperspectral image classification. *IEEE Transactions on Geoscience and Remote Sensing*, 55(5):2693–2705, 2017.

[134] Rajat Raina, Alexis Battle, Honglak Lee, Benjamin Packer, and Andrew Y Ng. Self-taught learning: transfer learning from unlabeled data. In *Proceedings of the 24th international conference on Machine learning*, pages 759–766, 2007.

[135] Rajat Raina. *Self-taught learning*. Stanford University, 2009.

[136] Zequn Jie, Yunchao Wei, Xiaojie Jin, Jiashi Feng, and Wei Liu. Deep self-taught learning for weakly supervised object localization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1377–1385, 2017.

[137] Grégoire Mesnil, Yann Dauphin, Xavier Glorot, Salah Rifai, Yoshua Bengio, Ian Goodfellow, Erick Lavoie, Xavier Muller, Guillaume Desjardins, David Warde-Farley, et al. Unsupervised and transfer learning challenge: a deep learning approach. In *Proceedings of the 2011 International Conference on Unsupervised and Transfer Learning workshop-Volume 27*, pages 97–111, 2011.

[138] Ravi K Samala, Heang-Ping Chan, Lubomir Hadjiiski, Mark A Helvie, Jun Wei, and Kenny Cha. Mass detection in digital breast tomosynthesis: Deep convolutional neural network with transfer learning from mammography. *Medical physics*, 43(12):6654–6666, 2016.

[139] Ravi K Samala, Heang-Ping Chan, Lubomir M Hadjiiski, Mark A Helvie, Caleb Richter, and Kenny Cha. Evolutionary pruning of transfer learned deep convolutional neural network for breast cancer diagnosis in digital breast tomosynthesis. *Physics in Medicine & Biology*, 63(9):095005, 2018.

[140] Jason Yosinski, Jeff Clune, Yoshua Bengio, and Hod Lipson. How transferable are features in deep neural networks? In *Advances in neural information processing systems*, pages 3320–3328, 2014.

[141] Zeynep Akkalyoncu Yilmaz. Cross-domain sentence modeling for relevance transfer with bert. Master's thesis, University of Waterloo, 2019.

[142] Mingsheng Long, Zhangjie Cao, Jianmin Wang, and Michael I Jordan. Conditional adversarial domain adaptation. In *Advances in Neural Information Processing Systems*, pages 1640–1650, 2018.

[143] Hana Ajakan, Pascal Germain, Hugo Larochelle, François Laviolette, and Mario Marchand. Domain-adversarial neural networks. *arXiv preprint arXiv:1412.4446*, 2014.

[144] Nour Eldeen M Khalifa, Mohamed Hamed N Taha, Aboul Ella Hassanien, and Sally Elghamrawy. Detection of coronavirus (covid-19) associated pneumonia based on generative adversarial networks and a fine-tuned deep transfer learning model using chest x-ray dataset. *arXiv preprint arXiv:2004.01184*, 2020.

[145] Edoardo Giacomello, Daniele Loiacono, and Luca Mainardi. Transfer brain mri tu-

mor segmentation models across modalities with adversarial networks. *arXiv preprint arXiv:1910.02717*, 2019.

[146] Shuyue Guan and Murray Loew. Using generative adversarial networks and transfer learning for breast cancer detection by convolutional neural networks. In *Medical Imaging 2019: Imaging Informatics for Healthcare, Research, and Applications*, volume 10954, page 109541C. International Society for Optics and Photonics, 2019.

[147] Yuxin Gong, Yingying Zhang, Haogang Zhu, Jing Lv, Qian Cheng, Hongjia Zhang Yihua He, and Shuliang Wang. Fetal congenital heart disease echocardiogram screening based on dgacnn: adversarial oneclass classification combined with video transfer learning. *IEEE transactions on medical imaging*, 2019.

[148] Divya Gopinath, Guy Katz, Corina S Pasareanu, and Clark Barrett. Deepsafe: A data-driven approach for checking adversarial robustness in neural networks. *arXiv preprint arXiv:1710.00486*, 2017.

[149] Weichen Zhang, Wanli Ouyang, Wen Li, and Dong Xu. Collaborative and adversarial network for unsupervised domain adaptation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 3801–3809, 2018.

[150] Ian Goodfellow. Nips 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*, 2016.

[151] HaiYing Wang, Rong Zhu, and Ping Ma. Optimal subsampling for large sample logistic regression. *Journal of the American Statistical Association*, 113(522):829–844, 2018.

[152] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[153] Haibo He and Edwardo A Garcia. Learning from imbalanced data. *IEEE Transactions on knowledge and data engineering*, 21(9):1263–1284, 2009.

[154] Rehan Akbani, Stephen Kwek, and Nathalie Japkowicz. Applying support vector machines to imbalanced datasets. In *European conference on machine learning*, pages 39–50. Springer, 2004.

[155] Hui Han, Wen-Yuan Wang, and Bing-Huan Mao. Borderline-smote: a new over-sampling method in imbalanced data sets learning. In *International conference on intelligent computing*, pages 878–887. Springer, 2005.

[156] Nitesh V Chawla, Aleksandar Lazarevic, Lawrence O Hall, and Kevin W Bowyer. Smoteboost: Improving prediction of the minority class in boosting. In *European conference on principles of data mining and knowledge discovery*, pages 107–119. Springer, 2003.

[157] Kung Jeng Wang and Angelia Melani Adrian. Breast cancer classification using hybrid synthetic minority over-sampling technique and artificial immune recognition system algorithm. *Int J Comput Sci Electron Eng (IJCSEE)*, 1(3):408–412, 2013.

[158] Kung-Jeng Wang, Bunjira Makond, Kun-Huang Chen, and Kung-Min Wang. A hybrid classifier combining smote with pso to estimate 5-year survivability of breast cancer patients. *Applied Soft Computing*, 20:15–24, 2014.

[159] Chris Seiffert, Taghi M Khoshgoftaar, Jason Van Hulse, and Amri Napolitano. Rusboost: A hybrid approach to alleviating class imbalance. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, 40(1):185–197, 2009.

[160] Tongan Cai, Hongliang He, and Wenyu Zhang. Breast cancer diagnosis using imbalanced learning and ensemble method. *Applied and Computational Mathematics*, 7(3):146–154, 2018.

[161] Mikel Galar, Alberto Fernandez, Edurne Barrenechea, Humberto Bustince, and Francisco Herrera. A review on ensembles for the class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 42(4):463–484, 2011.

[162] Qi Wang, ZhiHao Luo, JinCai Huang, YangHe Feng, and Zhong Liu. A novel ensemble method for imbalanced data learning: bagging of extrapolation-smote svm. *Computational intelligence and neuroscience*, 2017, 2017.

[163] Dennis H Murphree and Che Ngufor. Transfer learning for melanoma detection: Participation in isic 2017 skin lesion classification challenge. *arXiv preprint arXiv:1703.05235*, 2017.

[164] Irene Papanicolas, Liana R Woskie, and Ashish K Jha. Health care spending in the united states and other high-income countries. *Jama*, 319(10):1024–1039, 2018.

[165] Somaieh Goudarzvand, Jennifer St Sauver, Michelle M Mielke, Paul Y Takahashi, and Sunghwan Sohn. Analyzing early signals of older adult cognitive impairment in electronic health records. In *2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 1636–1640. IEEE, 2018.

[166] Somaieh Goudarzvand, Jennifer St Sauver, Michelle M Mielke, Paul Y Takahashi, Yugyung Lee, and Sunghwan Sohn. Early temporal characteristics of elderly patient cognitive impairment in electronic health records. *BMC medical informatics and decision making*, 19(4):149, 2019.

[167] Jui-Ting Huang, Jinyu Li, Dong Yu, Li Deng, and Yifan Gong. Cross-language knowledge transfer using multilingual deep neural network with shared hidden layers. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7304–7308. IEEE, 2013.

[168] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[169] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[170] Rebecca L Siegel, Kimberly D Miller, and Ahmedin Jemal. Cancer statistics, 2019. *CA: a cancer journal for clinicians*, 69(1):7–34, 2019.

[171] American Cancer Society. Cancer treatment and survivorship facts & figures 2014–2015. *Atlanta: American Cancer Society; 2014.*, 2014.

[172] Janet E Olson, Thomas A Sellers, Christopher G Scott, Beth A Schueler, Kathleen R Brandt, Daniel J Serie, Matthew R Jensen, Fang-Fang Wu, Marilyn J Morton, John J Heine, et al. The influence of mammogram acquisition on the mammographic density and breast cancer association in the mayo mammography health study cohort. *Breast Cancer Research*, 14(6):R147, 2012.

[173] Lynn C Hartmann, Thomas A Sellers, Marlene H Frost, Wilma L Lingle, Amy C Degnim, Karthik Ghosh, Robert A Vierkant, Shaun D Maloney, V Shane Pankratz, David W Hillman, et al. Benign breast disease and the risk of breast cancer. *New England Journal of Medicine*, 353(3):229–237, 2005.

[174] V Anuja Kumari and R Chitra. Classification of diabetes disease using support vector machine. *International Journal of Engineering Research and Applications*, 3(2):1797–1801, 2013.

[175] Arthur Asuncion. Uci machine learning repository, university of california,

irvine, school of information and computer sciences. *http://www. ics. uci. edu/˜ mlearn/MLRepository. html*, 2007.

[176] Igbe Tobore, Jingzhen Li, Liu Yuhang, Yousef Al-Handarish, Abhishek Kandwal, Zedong Nie, and Lei Wang. Deep learning intervention for health care challenges: Some biomedical domain considerations. *JMIR mHealth and uHealth*, 7(8):e11966, 2019.

[177] Nicolas Coudray, Paolo Santiago Ocampo, Theodore Sakellaropoulos, Navneet Narula, Matija Snuderl, David Fenyö, Andre L Moreira, Narges Razavian, and Aristotelis Tsirigos. Classification and mutation prediction from non–small cell lung cancer histopathology images using deep learning. *Nature medicine*, 24(10):1559–1567, 2018.

[178] Marylyn D Ritchie, Emily R Holzinger, Ruowang Li, Sarah A Pendergrass, and Dokyoon Kim. Methods of integrating data to uncover genotype–phenotype interactions. *Nature Reviews Genetics*, 16(2):85–97, 2015.

[179] Thomas A Lasko, Joshua C Denny, and Mia A Levy. Computational phenotype discovery using unsupervised feature learning over noisy, sparse, and irregular clinical data. *PloS one*, 8(6), 2013.

[180] Gloria Hyun-Jung Kwak and Pan Hui. Deephealth: Deep learning for health informatics. *arXiv preprint arXiv:1909.00384*, 2019.

[181] Karol Gregor, Ivo Danihelka, Andriy Mnih, Charles Blundell, and Daan Wierstra. Deep autoregressive networks. *arXiv preprint arXiv:1310.8499*, 2013.

[182] Partha Deb and Pravin K Trivedi. The structure of demand for health care: latent class versus two-part models. *Journal of health economics*, 21(4):601–625, 2002.

[183] Choong Ho Lee and Hyung-Jin Yoon. Medical big data: promise and challenges. *Kidney research and clinical practice*, 36(1):3, 2017.

[184] Lukasz A Kurgan and Petr Musilek. A survey of knowledge discovery and data mining process models. *The Knowledge Engineering Review*, 21(1):1–24, 2006.

[185] Byron C Wallace, Kevin Small, Carla E Brodley, and Thomas A Trikalinos. Class imbalance, redux. In *2011 IEEE 11th international conference on data mining*, pages 754–763. IEEE, 2011.

[186] Sotiris Kotsiantis, Dimitris Kanellopoulos, Panayiotis Pintelas, et al. Handling imbalanced datasets: A review. *GESTS International Transactions on Computer Science and Engineering*, 30(1):25–36, 2006.

[187] Victoria López, Alberto Fernández, Salvador García, Vasile Palade, and Francisco Herrera. An insight into classification with imbalanced data: Empirical results and current trends on using data intrinsic characteristics. *Information sciences*, 250:113–141, 2013.

[188] Gustavo EAPA Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD explorations newsletter*, 6(1):20–29, 2004.

[189] Qi Dong, Shaogang Gong, and Xiatian Zhu. Imbalanced deep learning by minority class incremental rectification. *IEEE transactions on pattern analysis and machine intelligence*, 41(6):1367–1381, 2018.

[190] Vicente García, José Salvador Sánchez, and Ramón Alberto Mollineda. On the effectiveness of preprocessing methods when dealing with different levels of class imbalance. *Knowledge-Based Systems*, 25(1):13–21, 2012.

[191] Mike Wasikowski and Xue-wen Chen. Combating the small sample class imbalance problem using feature selection. *IEEE Transactions on knowledge and data engineering*, 22(10):1388–1400, 2009.

[192] Social Sciences, Brian A Harris-Kojetin, Robert M Groves, Engineering National Academies of Sciences, Medicine, et al. Statistical methods for combining multiple data sources. In *Federal Statistics, Multiple Data Sources, and Privacy Protection: Next Steps*. National Academies Press (US), 2017.

[193] Nathaniel Schenker and Trivellore E Raghunathan. Combining information from multiple surveys to enhance estimation of measures of health. *Statistics in medicine*, 26(8):1802–1811, 2007.

[194] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3(1):9, 2016.

[195] Chongchong Yu, Rui Tian, Li Tan, and XY Tu. Integrated transfer learning algorithmic for unbalanced samples classification. *Acta Electronica Sinica*, 40(7):1358–1363, 2012.

[196] Zhixiang Yuan, Damang Bao, Zekai Chen, and Ming Liu. Integrated transfer learning algorithm using multi-source tradaboost for unbalanced samples classification. In *2017 International Conference on Computing Intelligence and Information System (CIIS)*, pages 188–195. IEEE, 2017.

[197] Aly A Mohamed, Wendie A Berg, Hong Peng, Yahong Luo, Rachel C Jankowitz, and Shandong Wu. A deep learning method for classifying mammographic breast density categories. *Medical physics*, 45(1):314–321, 2018.

[198] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[199] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(Mar):723–773, 2012.

[200] Chetak Kandaswamy, Luís M Silva, Luís A Alexandre, Ricardo Sousa, Jorge M Santos, and Joaquim Marques de Sá. Improving transfer learning accuracy by reusing stacked denoising autoencoders. In *2014 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 1380–1387. IEEE, 2014.

[201] Selen Uguroglu and Jaime Carbonell. Feature selection for transfer learning. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 430–442. Springer, 2011.

[202] Haitham Bou Ammar, Eric Eaton, Matthew E Taylor, Decebal Constantin Mocanu, Kurt Driessens, Gerhard Weiss, and Karl Tuyls. An automated measure of mdp similarity for transfer in reinforcement learning. In *Workshops at the Twenty-Eighth AAAI Conference on Artificial Intelligence*, 2014.

[203] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1–9, 2015.

[204] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[205] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[206] Bo Du, Wei Xiong, Jia Wu, Lefei Zhang, Liangpei Zhang, and Dacheng Tao. Stacked convolutional denoising auto-encoders for feature representation. *IEEE transactions on cybernetics*, 47(4):1017–1027, 2016.

[207] Etienne Laliberté and Pierre Legendre. A distance-based framework for measuring functional diversity from multiple traits. *Ecology*, 91(1):299–305, 2010.

[208] Syed Mumtaz Ali and Samuel D Silvey. A general class of coefficients of divergence of one distribution from another. *Journal of the Royal Statistical Society: Series B (Methodological)*, 28(1):131–142, 1966.

[209] I Csiszar. Eine information's theoretische ungleichung und ihre anwendung auf den beweis der ergodizitat von markoschen ketten, magyar tud. *Akad. Mat*, 1963.

[210] Tetsuzo Morimoto. Markov processes and the h-theorem. *Journal of the Physical Society of Japan*, 18(3):328–331, 1963.

[211] TM Cover and JA Thomas. Elements of information theory 2nd edn (hoboken, nj, john wiley & sons). 2006.

[212] Junier Oliva, Barnabás Póczos, and Jeff Schneider. Distribution to distribution regression. In *International Conference on Machine Learning*, pages 1049–1057, 2013.

[213] Inderjit S Dhillon, Subramanyam Mallela, and Rahul Kumar. A divisive information-theoretic feature clustering algorithm for text classification. *Journal of machine learning research*, 3(Mar):1265–1287, 2003.

[214] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan. An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures. In *null*, page 487. IEEE, 2003.

[215] Javier E Contreras-Reyes and Reinaldo B Arellano-Valle. Kullback–leibler divergence measure for multivariate skew-normal distributions. *Entropy*, 14(9):1606–1626, 2012.

[216] Gholamhossein Yari, Alireza Mirhabibi, and Abolfazl Saghafi. Estimation of the weibull parameters by kullback-leibler divergence of survival functions. *Appl. Math*, 7(1):187–192, 2013.

[217] Shun-ichi Amari and Andrzej Cichocki. Information geometry of divergence functions. *Bulletin of the polish academy of sciences. Technical sciences*, 58(1):183–195, 2010.

[218] Shili Lin. Kullback–leibler divergence for detection of rare haplotype common disease association. *European Journal of Human Genetics*, 23(11):1558–1565, 2015.

[219] M Vidyasagar. Bounds on the kullback-leibler divergence rate between hidden markov models. In *2007 46th IEEE Conference on Decision and Control*, pages 6160–6165. IEEE, 2007.

[220] Zhaolin Hu and L Jeff Hong. Kullback-leibler divergence constrained distributionally robust optimization. *Available at Optimization Online*, 2013.

[221] Arun Iyer, Saketha Nath, and Sunita Sarawagi. Maximum mean discrepancy for class ratio estimation: Convergence bounds and kernel selection. In *International Conference on Machine Learning*, pages 530–538, 2014.

[222] Dangna Li, Kun Yang, and Wing Hung Wong. Density estimation via discrepancy based adaptive sequential partition. In *Advances in neural information processing systems*, pages 1091–1099, 2016.

[223] Yan Li, Bhanukiran Vinzamuri, and Chandan K Reddy. Constrained elastic net based knowledge transfer for healthcare information exchange. *Data Mining and Knowledge Discovery*, 29(4):1094–1112, 2015.

[224] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin.

Liblinear: A library for large linear classification. *Journal of machine learning research*, 9(Aug):1871–1874, 2008.

[225] Samir Al-Stouhi and Chandan K Reddy. Transfer learning for class imbalance problems with inadequate data. *Knowledge and information systems*, 48(1):201–228, 2016.

[226] Min Zeng, Min Li, Zhihui Fei, Ying Yu, Yi Pan, and Jianxin Wang. Automatic icd-9 coding via deep transfer learning. *Neurocomputing*, 324:43–50, 2019.

[227] Priyanka Gupta, Pankaj Malhotra, Lovekesh Vig, and Gautam Shroff. Transfer learning for clinical time series analysis using recurrent neural networks. *arXiv preprint arXiv:1807.01705*, 2018.

[228] Yoshua Bengio. Deep learning of representations for unsupervised and transfer learning. In *Proceedings of ICML workshop on unsupervised and transfer learning*, pages 17–36, 2012.

[229] Sebastien Dubois, Nathanael Romano, Kenneth Jung, Nigam Shah, and David C Kale. The effectiveness of transfer learning in electronic health records data. 2017.

[230] Longfei Li, Yong Zhao, Dongmei Jiang, Yanning Zhang, Fengna Wang, Isabel Gonzalez, Enescu Valentin, and Hichem Sahli. Hybrid deep neural network–hidden markov model (dnn-hmm) based speech emotion recognition. In *2013 Humaine Association Conference on Affective Computing and Intelligent Interaction*, pages 312–317. IEEE, 2013.

[231] Djork-Arné Clevert, Thomas Unterthiner, and Sepp Hochreiter. Fast and accurate deep network learning by exponential linear units (elus). *arXiv preprint arXiv:1511.07289*, 2015.

[232] Antonio Gulli and Sujit Pal. *Deep Learning with Keras*. Packt Publishing Ltd, 2017.

[233] Ismail Babajide Mustapha and Faisal Saeed. Bioactive molecule prediction using extreme gradient boosting. *Molecules*, 21(8):983, 2016.

[234] SI Oke, MB Matadi, and SS Xulu. Cost-effectiveness analysis of optimal control strategies for breast cancer treatment with ketogenic diet. *Far East J Math Sci*, 109(2):303–342, 2018.

[235] John J Heine, Christopher G Scott, Thomas A Sellers, Kathleen R Brandt, Daniel J Serie, Fang-Fang Wu, Marilyn J Morton, Beth A Schueler, Fergus J Couch, Janet E Olson, et al. A novel automated mammographic density measure and breast cancer risk. *Journal of the National Cancer Institute*, 104(13):1028–1037, 2012.

[236] Steffen Moritz and Thomas Bartz-Beielstein. imputets: time series missing value imputation in r. *The R Journal*, 9(1):207–218, 2017.

[237] M Elter, R Schulz-Wendtland, and T Wittenberg. The prediction of breast cancer biopsy outcomes using two cad approaches that both emphasize an intelligible decision process. *Medical physics*, 34(11):4164–4172, 2007.

[238] Jason Brownlee. *Deep learning with Python: develop deep learning models on Theano and TensorFlow using Keras*. Machine Learning Mastery, 2016.

[239] McCoy R Scott C Farhadi A, Chen D, Vachon C miller J, and Ngufor C. Classification of breast cancer using deep transfer learning on structured healthcare data. *The 6th IEEE International Conference on Data Science and Advanced Analytics*, 2019.

[240] Robert A Jacobs. Increased rates of convergence through learning rate adaptation. *Neural networks*, 1(4):295–307, 1988.

[241] Sebastian Flennerhag, Pablo G Moreno, Neil D Lawrence, and Andreas Damianou.

Transferring knowledge across learning processes. *arXiv preprint arXiv:1812.01054*, 2018.

[242] Lindsay Haines and Charles L Shapiro. Cancer survivorship. *Mount Sinai Expert Guides: Oncology*, pages 535–542, 2019.

[243] Beata Strack, Jonathan P DeShazo, Chris Gennings, Juan L Olmo, Sebastian Ventura, Krzysztof J Cios, and John N Clore. Impact of hba1c measurement on hospital readmission rates: analysis of 70,000 clinical database patient records. *BioMed research international*, 2014, 2014.

[244] Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2011.

[245] Markus Ringnér. What is principal component analysis? *Nature biotechnology*, 26(3):303–304, 2008.

[246] Alina Beygelzimer, Sham Kakadet, John Langford, Sunil Arya, David Mount, and Shengqiao Li. Fnn: fast nearest neighbor search algorithms and applications. *R package version*, 1(1), 2013.

[247] Sylvain Boltz, Eric Debreuve, and Michel Barlaud. knn-based high-dimensional kullback-leibler distance for tracking. In *Eighth International Workshop on Image Analysis for Multimedia Interactive Services (WIAMIS'07)*, pages 16–16. IEEE, 2007.

[248] Sylvain Boltz, Eric Debreuve, and Michel Barlaud. High-dimensional statistical measure for region-of-interest tracking. *IEEE Transactions on Image Processing*, 18(6):1266–1283, 2009.

[249] Fuzhen Zhuang, Xiaohu Cheng, Ping Luo, Sinno Jialin Pan, and Qing He. Supervised

representation learning: Transfer learning with deep autoencoders. In *Twenty-Fourth International Joint Conference on Artificial Intelligence*, 2015.

[250] Francesco Mercaldo, Vittoria Nardone, and Antonella Santone. Diabetes mellitus affected patients classification and diagnosis through machine learning techniques. *Procedia computer science*, 112:2519–2528, 2017.