Medical Predictive Modelling using Transfer Learning

by

Samridhi Vaid

A thesis submitted in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science
University of Alberta

Abstract

Deep learning has had much success on challenging problems with large datasets. However, it struggles in cases with limited training data. Transfer learning represents a class of approaches for transferring knowledge from large source datasets to smaller target datasets. But many transfer learning approaches have constraints in terms of dataset size and similarity of output features. In this thesis, we introduce Quality-Diversity Transfer Learning (QDTL), a novel transfer learning approach based on neuroevolution for dealing with very small dataset problems with distinct output features. We demonstrate the success of QDTL on two medical prediction problems, outperforming standard transfer learning baselines.

Preface

This thesis presents an original work by Samridhi Vaid under the supervision of Dr. Matthew Guzdial. This work may be restructured to get published under different research venues in the near future.

Those living in an age of crisis must become pioneers of a better age, striving to find positive solutions and thereby turning the age into one of achievement.

-Arnold J. Toynbee

To my mentor Daisaku Ikeda and my parents Neena and Sumit

Acknowledgements

I would like to seize this opportunity to express my heartfelt gratitude to the following individuals who have played pivotal roles in the successful completion of this thesis:

First and foremost, I extend my deepest appreciation to my supervisor, Dr. Matthew Guzdial. Your guidance, expertise, and unwavering support have been invaluable throughout this research journey. Your insightful feedback, constructive criticism, and dedication to my academic and personal growth have played a pivotal role in shaping this thesis. I am truly grateful for the opportunity to work under your mentorship. Thank you for your patience, encouragement, and for constantly pushing me to achieve my best. I would also like to extend my thanks to my committee members Dr. Osmar Zaiane and Dr. Matthew Taylor for their valuable feedback, which helped me improve my thesis. Additionally, my gratitude goes to Dr. David Olson and Dr. Simon Urschel for providing me with the opportunity to work on their projects and for sharing valuable data.

I would like to express my sincere thanks to my parents, Sumit and Neena, for their unwavering love, encouragement, and belief in my abilities. Your constant support and sacrifices have been the foundation of my academic pursuits. I am grateful for your understanding, and motivation, and for always being there for me. To my sister Srishti and brother-in-law Ritik, I extend my heartfelt appreciation for your continuous support and understanding.

Deep gratitude goes to my friends, Mohan Sai Singamsetti, Kushankur Ghosh, Kushagra Chandak, Aakash Sasikumar, Subhojeet Pramanik, Deep Gandhi, Dhruv Mullick and Arghashree Banerjee. Your friendship, unwavering support and countless

discussions have played a crucial role in shaping my ideas and keeping me motivated during challenging times. Our shared experiences, including enduring -40 degrees Celsius, have created the most memorable moments of my life. I would also like to extend my thanks to my close friends Karishma, Natasha, Nidhi, Sukriti, Abhaas, Kelly and Parul for their unwavering support and friendship.

To all those mentioned above and to anyone else who has contributed in any way, big or small, to the completion of this thesis, I offer my heartfelt thanks. Your support, encouragement, and belief in me have been invaluable, and I am truly grateful for your presence in my life.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

Deep learning is currently used to address a wide variety of challenging problems. A large dataset is one of the key prerequisites for modern deep learning techniques to be effective [1]. As such, data limitations can serve as a major bottleneck to the application of deep neural networks (DNNs). Moreover, this problem may be more prevalent in the future, due to data pollution by large language models [2]. Regardless, currently a number of domains are missing out on benefitting from DNNs because they lack sufficient data. The medical domain, due to the costs and privacy concerns around data gathering, is one such domain.

Currently, low-data machine learning models like linear and logistic regressions are the most common models in the medical domain, despite their lower accuracy [3]. More complex techniques, like DNNs can be applied to low-data domains via transfer learning. Transfer learning has been applied to low-data problems in the medical domain [4, 5]. However, there are challenges associated with the application of transfer learning. Differences in data collection and variance in the output feature space between source and target domains can hinder the effectiveness of transfer learning methods [6]. Additionally, transfer learning models are prone to overfitting, especially when the target dataset is small, leading to poor generalization performance.

There exist transfer learning approaches like few-shot or zero-shot approaches to solve the above problems. They rely on large source datasets and are not applicable to our targeted tasks in the medical domain due to the limited dataset sizes and lack of shared classes. To overcome these limitations and enable the application of DNNs to more medical problems, we propose a novel transfer learning approach in this thesis called Quality-Diversity Transfer Learning (QDTL).

QDTL combines neuroevolution, architecture search, and conceptual expansion. Neuroevolution is a technique that applies evolutionary algorithms to optimize and evolve neural network architectures. We utilize quality-diversity (QD) optimization, which directly modifies the weights (parameters) of a DNN, enabling higher performance for low-training data tasks. QD optimization considers both the quality and diversity of solutions, which has the potential to mitigate overfitting. Architecture search allows us to optimize neural network architecture for the task at hand.

Furthermore, we incorporate the conceptual expansion representation proposed by Guzdial and Riedl [7], allowing us to describe the target model as a combination of weights from a source model. This has been helpful in prior works to help with low-data transfer learning [8, 9].

1.1 Research Questions and Related Contributions

In this thesis, we try to overcome the low data limitation problem for the medical prediction task. To overcome this limitation, we attempt to answer the following research questions:

1. Can we outperform standard logistic regression in medical prediction tasks with our approach?

2. Can we outperform existing transfer learning methods by leveraging a search-based transfer learning approach for predictive care in the medical domain?

3. Is Quality-Diversity the optimal choice among search-based optimization techniques?

The following are the contributions of this thesis:

1. A novel transfer learning approach to solve low-data problems via quality-diversity-based neuroevolution: Quality-Diversity Transfer Learning (QDTL)
2. The application of QDTL to the task of predicting pre-term birth, with the ability to predict births within 7 days, a massive improvement over the current state of the art.
3. The application of QDTL to predict the survival days of patients undergoing organ transplantation surpassing the finetuning approach traditionally applied to transfer learning medical prediction problems.

1.2 Thesis Outline

This thesis is organized into four chapters, including the introduction. In Chapter 2, the necessary background knowledge is presented to understand the concepts employed in this thesis. Chapter 3 focuses on our proposed approach, Quality-Diversity Transfer Learning (QDTL), and its application to two medical domain problems. The first task involves predicting the gestational age of pregnant women to anticipate preterm birth (PTB). The second task entails conducting a survival analysis of patients who have undergone organ transplants and predicting the post-transplant survival time until the final check-up. Detailed experiments and comparisons with baselines are provided in this chapter, emphasizing the performance of QDTL in comparison to different baselines. Finally, chapter 4 concludes the thesis by summarizing the findings, discussing future work, and acknowledging the limitations encountered.

Chapter 2

Background

This chapter serves as a resource providing information on various topics crucial for readers to understand the content presented in this thesis. Sections 2.1, 2.2, 2.3, and 2.4 provide a brief overview of Artificial Neural Networks, Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM) Recurrent Neural Networks, and Bidirectional Long Short-Term Memory (BiLSTM) Recurrent Neural Networks. LSTMs and BiLSTMs were employed as the primary models in this thesis.

Section 2.5 presents the concept of transfer learning, highlighting its significance in our research as we introduce a novel transfer learning approach. Additionally, the chapter covers combinational creativity, Conceptual Expansion based Monte Carlo Tree Search (CE-MCTS), evolutionary search and architecture search forming the foundation of our proposed approach known as Quality-Diversity Transfer Learning (QDTL).

To showcase the effectiveness of our approach, we explore two application domains: Preterm Birth (PTB) and Organ Transplant and we cover related work in these two domains.

2.1 Artificial Neural Networks

Artificial neural networks (ANNs) are computational models loosely inspired by the human brain. This section will delve into the fundamental concepts and principles underlying artificial neural networks, providing an understanding of their architecture, learning algorithms, and applications. We have shown a simple ANN in Figure 2.1

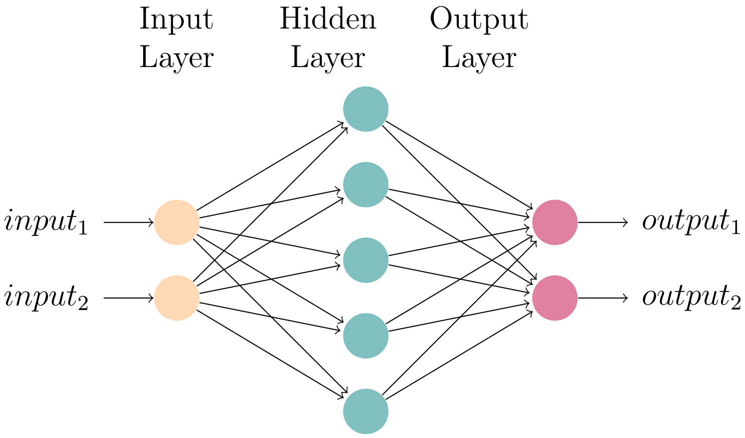

Figure 2.1: A Simple Artificial Neural Network

Artificial neurons, also known as perceptrons, form the fundamental units of an artificial neural network. They are designed to mimic the behaviour of biological neurons. Each artificial neuron receives inputs, applies weights to those inputs, and produces an output using an activation function. The activation function determines the neuron's response based on the sum of weighted inputs as shown in Equation 2.1.

$$Y = Activationfunction(\sum(weights * input + bias)) \tag{2.1}$$

Activation functions introduce non-linear transformations to the weighted sum of inputs in artificial neurons. Some common activation functions are the sigmoid function, hyperbolic tangent function, and rectified linear unit (ReLU) [10]. The choice of activation function depends on the specific problem and network architecture.

An artificial neural network consists of multiple interconnected layers of perceptrons. The basic architecture is composed of an input layer, one or more hidden layers, and an output layer. The input layer takes in features, which are then processed by the hidden layers, and passed on to the final layer that computes the desired output.

Feedforward neural networks are a type of ANN in which information flows in only one direction, from the input layer to the output layer, without any loops. In the training process, the network's weights are updated to minimize the difference between predicted outputs and actual outputs, which is known as loss. Backpropagation, a widely employed algorithm for training neural networks, calculates the gradients which is a partial derivative of the loss function with respect to the weights to minimize this loss. Optimization algorithms like gradient descent are utilized to change the weights of the network so as to minimize the loss.

Overfitting occurs when a neural network performs well on the training data but not on the test data. Regularization techniques are employed to mitigate overfitting. Common approaches include L1 and L2 regularization, which introduce penalties for large weight magnitudes, and dropout, which randomly deactivates a fraction of neurons during training. Using our approach proposed in Chapter 3, we try to solve the overfitting problem for domains with low-data problems.

2.2 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a type of artificial neural network which processes sequential data. Unlike traditional feedforward neural networks, which process inputs independently, RNNs have a form of memory that enables them to maintain an internal state or hidden representation. This hidden state is updated recurrently as the network processes each input in the sequence, allowing the network to retain information about the sequence's context and dependencies. Figure 2.2 shows a simple RNN architecture.

Traditional RNNs tend to have limited short-term memory, making it challenging to capture long-range dependencies in sequences. The recurrent connections that carry information over time may degrade or vanish over long sequences, resulting in the network's inability to effectively utilize past information.

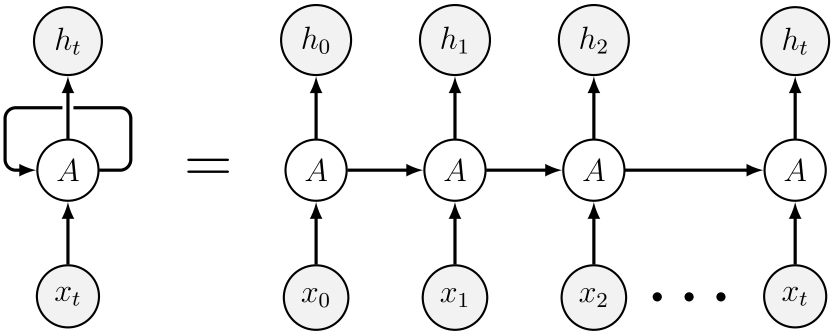

Figure 2.2: Simple Recurrent Neural Network (RNN) Diagram

2.3 Long Short-Term Memory

LSTMs were specifically designed to address the limitations of traditional RNNs in capturing long-term dependencies in sequential data. LSTMs process long-term dependencies because of their ability to selectively retain information.

We cover the LSTM components as we use LSTMs for both problem domains discussed in Chapter 3. The cell state is regulated by three main components: the input gate, the forget gate, and the output gate.

- Input Gate: The input gate decides which information is allowed in the cell state. It takes the current input and the previous hidden state as input and applies a sigmoid activation function to output a value between 0 and 1 for each element in the cell state.

- Forget Gate: The forget gate decides which information to discard from the cell state. It takes the current input and the previous hidden state as input and applies a sigmoid activation function to output a value between 0 and 1 for each element in the cell state.

- Output Gate: The output gate determines the amount of information to be output from the cell state. It takes the current input and the previous hidden state as input and applies a sigmoid activation function to output a value between 0 and 1 for each element in the cell state. It also applies a tanh activation function to squish the values between -1 and 1.

The LSTM cell combines the outputs of these gates to update the cell state and produce the hidden state at each time step. The hidden state carries the processed information from the input sequence and is used as the output of the LSTM layer or as input for subsequent layers in the network.

LSTMs have proven to be effective in capturing and modelling long-term dependencies in sequential data. Their ability to selectively retain or forget information over time allows them to mitigate the vanishing and exploding gradient problems commonly encountered in traditional RNNs. This enables LSTMs to excel in tasks that involve processing sequences with complex dependencies.

2.4 Bidirectional Long Short-Term Memory

Bidirectional LSTM (BiLSTM) is an extension of the Long Short-Term Memory (LSTM) network. BiLSTM process sequential data in both forward and backward directions simultaneously. This allows them to have access to past and future information at each time step.

BiLSTM has two separate LSTM layers. The forward LSTM layer processes the input sequence from the beginning to the end, while the backward LSTM layer processes it in the reverse direction. The use of BiLSTM is particularly beneficial in tasks where the prediction at each time step depends not only on past observations but also on future context [11]. We use BiLSTM in our approach where we had more data to figure out large-range dependencies.

2.5 Transfer Learning

Transfer learning is an approach that adapts existing knowledge from one domain to another. It involves transferring knowledge learned from a source domain to a target domain. Transfer Learning approaches can be categorized into four types: instances-based, mapping-based, network-based, and adversarial-based [12]. Most of the prior work in transfer learning makes use of the finetuning approach [13]. Finetuning is the process in which a pre-trained model's weights are further adjusted using a smaller, task-specific dataset. In this thesis, we use the finetuned model as the base model for our approach described in chapter 3.

Below are the key components and steps involved in transfer learning:

- Source Task and Pretrained Model: Transfer learning starts with a source task, where a model is trained on a large dataset. The pretrained model, often a deep neural network, learns to extract useful features and representations from the source data.

- Target Task and Target Data: The target task is the task for which we want to apply transfer learning. In most cases, the source and target tasks are from different domains. It sometimes might be the case that the target task has limited labelled data.

- Feature Transfer: In transfer learning, we transfer the features from the source domain to the target domain. A common approach to do this is finetuning [13].

- Transfer Learning Evaluation: The performance of the transfer learning approach is evaluated on the target task using appropriate metrics such as accuracy, precision, recall, or F1 score. By leveraging the knowledge from the source task, transfer learning aims to improve the performance of the model on the target task, especially when the target data is limited.

We discuss transfer learning, as our approach described in Chapter 3 is a novel transfer learning method to resolve the issue of data scarcity.

2.6 Combinational Creativity

Combinational creativity, also known as combinatorial creativity, refers to the generation of novel ideas by combining existing concepts, elements, or components [14]. Combinational creativity (CE) has been used in previous work. Guzdial and Ried [7] introduced an approach called conceptual expansion. In conceptual expansion, a particular weight is represented as a combination of α and *f*. Representing a particular weight in this manner allows us to discover combinations of weights [9]. Equation 2.2 represents CE, where the weights of the DNN model serve as the components that are recombined:

$$CE(F, \alpha) = \alpha_1 * f_1 + \alpha_2 * f_2 + ... + \alpha_n * f_n \tag{2.2}$$

Where W represents a weight in the final output model, F = $f1, f2, ..., fn$ represents existing weights and $\alpha = \alpha_1, \alpha_2, ..., \alpha_n$ are alpha value filters, which undergo pairwise multiplication with the weights. The alpha filters play a vital role in modifying the weights during the combination process. Within the conceptual expansion (CE) framework, the same f value can be present multiple times in F but with different alpha (α) values. This flexibility allows CE to represent a diverse set of combinations [8] We have used conceptual expansion for our approach which allows us to describe the target model as a combination of the weights from a source model, modelling the source model knowledge as a combinational creativity task.

2.7 Conceptual Expansion-based Monte Carlo Tree Search (CE-MCTS)

Conceptual Expansion-based Monte Carlo Tree Search (CE-MCTS)[8] is a prior work most related to our work and one of the approaches we compared our proposed approach with. Mahajan and Guzdial used a tree-based transfer learning approach to learn the behaviour of an individual on a target task using the data of that same individual available on a secondary task in the financial domain. They proposed a solution called

Conceptual Expansion-based Monte Carlo Tree Search (CE-MCTS) based on transfer learning and combinational creativity and showed its effectiveness in limited data scenarios. The approach involves training a source model on other individuals' data for the target task and using data from the specific individual on the secondary task to guide MCTS in the search for weight combinations to approximate a final model. CE defines the search space of possible models by representing combinations of existing knowledge (weights from the source model). Similar to their approach, we are using a finetuned model as an input. The major differences between their work and ours are that they employ Monte Carlo Tree Search (MCTS) instead of Quality-Diversity Search and that they did not employ architecture search, meaning they searched over fixed models. We anticipate that our approach will outperform MCTS because of these two differences because Quality-Diversity Search uses a population of points and the diversity factor allows us to find far-off models in the search space.

2.8 Evolutionary Search

Evolutionary algorithms are a class of computational techniques inspired by the process of biological evolution. They are used to solve optimization problems by mimicking natural selection and genetic mechanisms [15].

In the evolutionary algorithm, a population of candidate solutions is generated. Using a process called a mutation, with some probability, the potential solution is replaced with its variation. Then using a process called crossover a new population is derived by combining solutions in a domain-dependent way. This new population is added to the existing population. The reduction process helps in converging the population by gradually eliminating less fit individuals to select the top K solution using some fitness function where K is the population size [15].

Quality diversity algorithms are a specific subset of evolutionary algorithms that emphasize the generation of a diverse set of high-quality solutions. Traditional evolutionary algorithms often focus solely on optimizing a single objective or finding the single best solution. In contrast, quality diversity algorithms aim to explore and maintain a diverse set of solutions that cover different aspects of the problem space.

Quality diversity algorithms utilize two fitness functions that take into account both the quality and diversity of solutions. Diversity represents how different a solution is compared to the existing solutions in the population. By promoting diversity, these algorithms can uncover a wide range of solutions, allowing for a more comprehensive exploration of the problem space. These algorithms have been applied in various fields, including robotics, optimization, game design, and generative art, to name a few, where exploring and discovering a range of solutions is valuable [16–18].

2.9 Architecture Search

Architecture search is a method used in machine learning to automatically discover the optimal neural network architecture for a given task [19]. The roots of this approach can be traced back to the 1980s when researchers introduced evolutionary optimization techniques aiming to discover both the architectures and weights of neural networks [20].

Singamsett et al. [9] demonstrated successful results by employing evolutionary-based search methods combined with transfer learning in the domain of image classification. In our work, we have introduced a unique approach that combines Quality-Diversity (QD) and evolutionary-based search. While our approach shares similarities with the previous work, there are notable differences in terms of the optimization approach, the emphasis on regression tasks, and the specific setup of the architecture search problem. Our GA ablation can be considered as an approximation of their approach, albeit with some distinct characteristics. Liu et al. [21] combined federated learning with neural architectures search to do image classification for providing medical data security solutions. Our work we are not employing federated learning. Guo et al. [22] used a stochastic deep collection method that used neural architecture search to help protect the groundwater quality. Our work differs in optimization approach, regression task emphasis, and architecture search setup. We don't compare our approach with these works as our contribution is not to architecture search as a field.

2.10 Preterm Birth

Preterm birth is defined as the delivery of a baby before 37 full weeks of gestation. In a full-term pregnancy, the gestational period typically lasts around 40 weeks. Preterm birth is the leading health problem during the perinatal period, the leading cause of death in children < 5, and a major cause of chronic disease [23–25]. A diagnostic approach for predicting preterm delivery is the need of the hour as it will help communities worldwide and will stimulate the creation of treatments against preterm birth. Many clinical laboratory tests are also expensive and therefore not attainable for those populations most at risk for PTB [26–28]. Moreover, the existing DNN approaches would require a lot more data than what is available for PTB due to data availability and privacy concerns.

2.11 Organ Transplant

Organ transplantation is a medical procedure in which a failing or damaged organ is replaced with a healthy organ from a donor [29]. It allows people with terminal organ failure and no other treatment options to receive a donor organ from another human. Commonly transplanted organs include the heart, kidneys, liver, lungs, and pancreas [1]. Days to the endpoint measure the number of days after an organ transplant the patient survives censored at the end of the observation period. Censored at the end of the observation period means that the survival time for some patients may not be fully observed or known because the study or observation period has ended before the event (e.g., death) occurs for those patients.

[1] https://optn.transplant.hrsa.gov/data/view-data-reports/national-data/

Chapter 3

Quality-Diversity Transfer Learning (QDTL)

In this chapter, we delve into an in-depth explanation of our approach known as Quality-Diversity Transfer Learning (QDTL). The chapter begins with Section 3.1, providing an overview of the QDTL approach. It encompasses a comprehensive understanding of the problem definition and the architectural framework. Section 3.2 is dedicated to presenting and describing the two specific domains we have chosen for evaluation and experimentation. This section provides insights into the characteristics and peculiarities of these domains, setting the context for the subsequent analysis. The evaluation of our approach is thoroughly discussed in Section 3.3, where we provide an in-depth examination of the experimental setup, methodologies, and metrics used to assess the performance of QDTL.

3.1 System Overview

In this section, we present our Quality-Diversity Transfer Learning (QDTL) approach[1]. Our approach is based on domain transfer, where we have distinct source and target domains and transfer knowledge from the source to the target domain. Our approach can be defined as a three-step process. We first train a model on the source domain data, this step is domain-specific. We then use the weights of this model trained on

[1] https://github.com/SamridhiVaid/Medical-Predictive-Modelling-using-Transfer-Learning

source domain data and further finetune it on target domain data. Finally, we take the model finetuned on the target domain data and run quality-diversity over the model weights to find the most optimal model to solve our problem. To the best of our knowledge, this is the first time QD has been applied as neuroevolution for supervised transfer learning.

3.1.1 Problem Definition

In this thesis, we focus on low-data medical prediction tasks. By low-data, we mean a <200 instance dataset. These medical domain problems are structured such that we take as input a series of features representing an individual patient and output a prediction in terms of the time until some future medical event. We assume we have access to two datasets. A small dataset that represents our target task with less than 140 instances and a larger dataset with 1000 instances represents another medical prediction task. The problem then is how best to adapt the knowledge from the larger source domain dataset to the smaller target domain dataset. Most of the existing machine learning approaches would not work in our case because firstly, our dataset is very small invalidating most non-transfer learning approaches. Secondly, our source task and target task have some shared input features but the output feature is different for our source and target task, invalidating many transfer learning approaches.

3.1.2 Architecture

We use a bidirectional Long Short Term Memory (Bi-LSTM) Recurrent Neural Network-based architecture as our source model. Consisting of 4 Bi-LSTM layers. Each of the 4 Bi-LSTM layers contains 512 units with a dropout size of 0.2. They use the default linear activation function. These 4 Bi-LSTM layers are followed by a dense prediction layer using the default linear activation function at the end with a unit size of 1. We employed Keras and used all the default values for its Bi-LSTM and Dense layers otherwise. We use a small model because of the lack of training data and to avoid overfitting. We use RMSprop as our optimizer and mean square error (MSE) for calculating loss for both step one where we train our model on the source data and step two, where we finetune the model on our target data. We call the model from step one as the source model and from step two as the target model. Since we are predicting numerical values, we chose MSE as our loss function. For our target model, which is the basis for our approach and all baselines and ablations, we use all the same parameters as our source model but instead of 4 Bi-LSTM layers, we use 4 LSTM layers. We used the bidirectional LSTM model for our source model because it has a larger dataset size, moreover, the data is not sequential. The Bi-LSTM model works better with larger datasets and helps us figure out larger-range dependencies in the case of the source model. For our target model, we used an LSTM model as this allowed us to compare to CE-MCTS [8] more easily.

3.1.3 Quality-Diversity Transfer Learning

Quality-Diversity (QD) is an evolutionary algorithm, which involves generating and evolving a population of models over many generations. Our QDTL algorithm uses QD to optimize the performance of a DNN model trained on a source task for our target task. We employed QD over other optimization approaches in pursuit of a variety of models to reduce the risk of overfitting. We hypothesized this would allow for higher performance of our final DNN models for our low-data tasks [19]. In our approach, the fitness of each generated model is evaluated based on quality and diversity fitness objectives, and these selection pressures encourage both high-quality and diverse solutions in the population. Our approach assumes two objective functions $f_Q : R^n \to R$ and $f_D : R^n \to R$ for quality and for the diversity populations respectively. For each model $s_Q \epsilon S_Q$ and $s_D \epsilon S_D$, where S_Q is our generated quality population and S_D is our diversity population, the goal of our approach is to select models from these population based on their respective objective functions and then select the top 10 models based the quality fitness function. We finally return the single best model.

Algorithm 1 represents our entire approach. We first train our model on the source dataset, which in our case is a Cervical Cancer dataset[2] with 858 samples. We chose this particular dataset as our source dataset due to the domain similarity and some feature similarity with our target domains. For example, demographic information and historical medical data are some similar features. This dataset has 46 input features. We employed a 60/40 train-validation split, we used this split to make the size of the training set closer to the target dataset across both domains. We train the neural network architecture described above for 100 epochs, with a learning rate of 0.0001. We determined all hyperparameter values based on the validation set. We further finetune this model on training data from our target task. This is our baseline model, which we refer to as finetune. This target model is given as input to a quality-diversity process, in which we optimize the weights directly according to the target training dataset. Line 1 represents this input model.

We run our QD architecture search and transfer process to output the final target model. Based on the input, we initialize two populations (i.e., quality population and diversity population) of fixed size by running our mutation function $popsize - 1$ times. As with other evolutionary algorithms, QD requires specialized mutation, crossover, and fitness functions, which are described in the subsections below. We run the search process for 20 generations, then select the top 10 models according to our quality fitness and return the final best model. We chose 20 generations and our other hyperparameters based on the validation performance of our first domain.

3.1.4 Crossover

In this subsection, we discuss our QDTL crossover function. In quality-diversity search, the crossover function combines genetic information from two or more parent models to create a new offspring model that explores different regions of the search

[2] https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification

Algorithm 1 QDTL Approach

```
Input: An architecture $A$, the population size $pop\_size$, maximal generations $gen$,
       the $source$ dataset, and the $target$ dataset.
Output: Best performing architecture.
$A \leftarrow$ train $A$ on $source$
$pop_Q = \{A\ \}$
$pop_D = \{A\ \}$
while $|pop_Q| < pop\_size$ do
    $network_Q \leftarrow$ Mutation($A$)
    $network_D \leftarrow$ Mutation($A$)
    $pop_Q$.append($network_Q$)
    $pop_D$.append($network_D$)
end
i $\leftarrow$ 0
while $i < gen$ do
    $pop \leftarrow$ Crossover($pop_Q$, $pop_D$)
    $pop \leftarrow$ Mutation($pop_Q$, $pop_D$, $mutationRate$)
    $pop \leftarrow$ Reduce($pop$, $fitness\_Q$, $fitness\_D$)
    i $\leftarrow$ i + 1
end
architecture = best_model($pop_Q$)
return architecture
```

space and promotes diversity among the models. We conduct crossover between both populations. First, we select two parents via a weighted sampling based on the quality and diversity of fitness scores. Second, we identify the position of the LSTM layers in both selected parents. Then we randomly choose one position in the LSTM layers of both parents. We only track LSTM layers as these are the layers in which feature extraction occurs. Using this position, we take the initial half of the weights and layers from the first model, and the latter half of the weights and layers from the second model. By creating new models from the quality and the diversity population, we hope to combine the good qualities of existing models and promote diversity in the search space.

3.1.5 Mutation

We have eleven mutation functions and randomly select one of them for our mutation process. Conceptual expansion is utilized in our mutation functions, which involves broadening the scope of a concept beyond its original meaning, in our case that is the model weights. Conceptual expansion re-represents a neural network as a combination of α and f values, where f represents the features (weights) of some input model and α represents a matrix paired to each f value. To calculate the final weight we have each paired α and f value undergo pairwise multiplication then sum across all pairs. For further detail please see the original paper [7].

By representing a specific weight as a combination of α and f, we are able to discover high-quality weight combinations. Directly modifying the α and f values associated with each weight allows us to manipulate the network weights [8]. Our first mutation four functions are the ones used for conceptual expansion in the work by Guzdial and Reild [7]. These were successfully employed by Mahajan's transfer learning approach [8]. As a result, we use the same four functions in our approach. For our remaining seven functions, we take inspiration from the work by Singamsetti et al. [9] and in addition to using two of their mutation functions which are, adding random α and f values to a random position to a random mutation layer, we extend these by using subtraction, multiplication and division to a random layer and index. Below are the mutation functions we used in this work

- The first function multiplies a randomly selected index of a randomly selected α with a scalar value in the range [-2, 2].
- The second function multiplies an entire randomly selected α matrix by a scalar value in the range [-2, 2].
- The third function swaps two randomly chosen α and f values (with equivalent dimensions).
- The fourth function adds two randomly chosen α and f values (with equivalent dimensions) to a CE approximation.

- The fifth function adds a random α and f to a random position to a random mutation index
- The sixth function adds a random α and f to a random position to a random mutation layer
- The seventh function subtracts a random α and f from a random mutation index
- The eight function subtracts a random α and f from a random mutation layer
- The ninth function multiplies a random α and f to a random mutation index
- The tenth function multiplies a random α and f to a random mutation layer
- The eleventh function divides a random α and f from a random mutation index

3.1.6 Fitness Score

We use two fitness scores, one for the quality population and the other for the diversity population. For the quality fitness function, we evaluate the mean square error on the training set. For the diversity population, we use exploratory fitness. For this, we create a matrix of the weights of the input model which is represented W_B and take the absolute difference with the matrix of weights of the child model which is represented by W_C. We return the mean of this difference. We use this exploratory fitness function to find models in the search space that are far apart from our input model.

3.2 Domain

We test our algorithm for two tasks in the medical domain. For our first task, we predict the gestational age of pregnant women in order to predict preterm birth (PTB). Gestational age is the number of days after which the woman will deliver. For our second task, we conduct a survival analysis of patients who have undergone organ transplants and we predict the number of days after the transplant that these patients survive (up to a final check-up). Since both delivery date prediction and survival days after an organ transplant are more complex functions that cannot be reduced to a simple linear or logistic function, we believe DNNs are better suited to solve this problem. To demonstrate this, we performed logistic regression on one of our tasks and compared its performance with that of finetuning. It surpassed logistic regression. One of the most basic requirements of modern DNNs is a large dataset. However, there are limited samples in both cases, and a naive application of DNNs would not provide an effective solution. We, therefore, hypothesized that a transfer learning approach would be better suited to solve these problems. It is important to note that we obtained informed consent to utilize this data for our research purposes.

3.2.1 Preterm Birth (PTB)

PTB is defined as birth before 37 full weeks of gestation. Preterm birth is the leading health problem during the perinatal period, the leading cause of death in children less than 5 years of age, and a major cause of chronic disease [23–25]. A diagnostic approach for predicting preterm delivery is the need of the hour as it will help communities worldwide and will stimulate the creation of treatments against preterm birth. Many clinical laboratory tests are also expensive and therefore not attainable for those populations most at risk for PTB [26–28]. Moreover, the existing DNN approaches would require a lot more data than what is available for PTB due to data availability and privacy concerns. Our approach not only tries to overcome the low-data limitation problem but predicts women at risk for a PTB, and actually predicts when these women will deliver. This output would be helpful in designing appropriate treatment for these women.

For this problem, we had 70 features and 135 total samples. We got these 135 samples from Christiaens et al.'s work [30]. Where 43 samples were of women who had PTB and the remaining 92 samples of women who had normal delivery or non-preterm birth (NPTB) samples. Since we had two sets of data PTB and NPTB, for this task we finetune the source model trained on the cervical cancer dataset twice. First, we finetune the model on an NPTB dataset and then further finetune on a combination of PTB and NPTB samples for 30 epochs each with a learning rate of 0.0001. This model then serves as the input to our QD approach.

3.2.2 Organ Transplant

Solid organ transplantation allows people with terminal organ failure and no other treatment options to receive a donor organ from another human. In our case, samples of children requiring heart, kidney or liver transplants have been collected in a national collaboration before and 3 and 12 months after transplantation. We predict days to the endpoint, which is the number of days after an organ transplant the patient survives censored at the end of the observation period. For the organ transplant problem, we had 130 samples in this task and 158 features. For this task, since we have a single dataset, we only finetune the source model which is trained on the cervical cancer dataset once for 45 epochs, with a learning rate of 0.001.

3.3 Evaluation

In this thesis, we try to overcome the low-data limitation problem in the medical domain by utilizing transfer learning and quality-diversity optimization. We show how well our approach performs on two tasks from the medical domain by evaluating our approach on five-fold cross-validation. Since for both tasks, we had limited data, moreover due to the nature of the domain, a small improvement in performance can have a significant impact on patient outcomes. Hence, we employed a five-fold cross-validation.

For our first task, the Preterm Birth (PTB) we are predicting the gestational age. Since we had PTB and NPTB data on which we were finetuning our base model, we created minifolds in addition to five-fold cross-validation. We wanted to test out all combinations of the two datasets. For example, for fold1 we use 80% of NPTB data for training the model. We represent this 80% data as A, B, C and D where each represents 20% NPTB data. We interchangeably choose 3 folds out of these and use them for the initial finetuning of the source model. The remaining one fold is combined with 80% PTB data and used for the final finetuning of the model we get from the previous step. We repeat this step 4 times, each time combining one of the NPTB folds with 80% PTB data. We further perform this for each of the remaining 4 folds so as to make sure our model is robust. For our second task, the organ transplant task we did a simple 5-fold cross-validation, with an 80/20 train-test split.

We used fixed seeds for all tests and computed the average mean square error and standard deviation on the test dataset to evaluate the performance of different approaches. All of the baseline approaches, ablations and our QDTL approach are executed roughly for around 4 hours independently. These experiments are conducted using 6 CPUs and 2 NVIDIA Tesla V100 GPUs per task.

3.3.1 Baselines and Ablations

In this thesis, we employed two baselines and created four ablations. For all the ablations we use the same eleven mutation functions as QDTL and we run each for the same 20 iterations for a fair comparison. For the organ transplant task, we only use the two baselines for comparison as the PTB task already shows the relative performance of the ablations. For the baselines, it was essential to compare QDTL with a finetuning baseline given its prevalence. We additionally show the performance on a no-transfer learning baseline as this is the more common approach.

Baselines

- The first baseline approach, which we refer to as No-Transfer Learning, involves directly training the model on the training dataset without any transfer learning. For the PTB task, we trained the model for 30 epochs, while for the Organ

Transplant task, we trained it for 45 epochs. We used learning rates of 0.0001 and 0.001, respectively. We selected these hyperparameters based on preliminary experiments conducted to optimize the training process.

- The second baseline, called Finetune, involves finetuning our source model on the target domain data. This represents the standard approach one might take to solve this type of transfer learning problem for a medical prediction task [31, 32]. We utilized this model as an input to all other baselines, ablations, and our QDTL approach, making it an additional ablation of our approach.

Ablations

- The first ablation is called CE-MCTS and is adapted directly from Mahajan and Guzdial [8]. For this ablation, we removed the architecture search. In addition, MCTS is used over QD without the diversity fitness score. We used the same setup but executed their approach with 20 iterations of 10 rollouts of length 10.

- The second ablation, referred to as "Random Walk," employs a random walk to explore the model space. For this ablation, we have removed the architecture search portion of our approach. In addition, a random walk can be understood as equivalent to QD without a population and without a crossover function. This allows us to test whether the mutation functions alone are sufficient to achieve high-quality results. In this technique, at each step, a random child is selected, and this process is repeated for 20 iterations. The selection of the best model is based on the quality fitness function criteria used in our QDTL approach, facilitating a straightforward comparison between the techniques.

- The third ablation, referred to as "Greedy," employs a greedy search strategy to explore the model space. For this ablation also we have removed the architecture search portion of our approach. In addition, a greedy search can be understood as equivalent to QD without a population and without a crossover function. In this technique, the mutation function described earlier is utilized to generate ten random neighbours. The neighbour with the highest node value is selected at each step, and this process is iterated for 20 steps. The best final model is chosen based on the quality fitness function criteria used in our QDTL approach.

- The fourth ablation, known as GA, employs a genetic algorithm with the same configuration as our QDTL approach but we remove the diversity population and diversity fitness score. The purpose of including this ablation was to demonstrate that the dual optimization in QD allows for a more beneficial exploration of the search space.

3.3.2 Comparison with Logistic Regression

As a preliminary experiment, we compared logistic regression to finetuning on the Preterm Birth (PTB) task on the basis of the number of days by which the model was off by or the error in days. We only compared finetuning to Logistic Regression and did not compare it with other low-data machine learning techniques like Decision Trees or Support Vector Machines because Logistic Regression is the most common approach in medical domain tasks [33]. We expected SVMs and Decision Trees to perform similarly to logistic regression due to the complex nature of our problem which is difficult to solve using these linear models. Moreover, ours is a regression task and all these models are more compatible for classification tasks.

For these experiments, we employed a slightly modified setup since they were conducted as part of the initial configuration for our research. For the finetuning model, we trained the model on 80% of the cervical cancer dataset[3], This is our source model. We used an 80/20 train-test split for this. We finetuned this model on 80% non-preterm birth data (normal pregnancy dataset). Finally, we finetuned this model on 20% non-preterm birth data and 20% PTB data. We evaluated the performance of our model on the remaining 80% PTB data.

[3] https://www.kaggle.com/datasets/loveall/cervical-cancer-risk-classification

For our logistic regression model, to compare its performance to the finetuned model we trained it on 20% non-preterm birth data and 20% PTB data. We evaluated the performance of our model on the remaining 80% PTB data. Our results are shown in Table 3.1. These results clearly show the superiority of finetuning, which our approach outperforms as we cover below. Thus we do not include a logistic regression baseline for the other experiments. These results also address our first research question.

Table 3.1: Logistic Regression vs. Finetuned performance in terms of the average error in days

Approach	Error in Days
Logistic Regression	20.5±18.19
Finetune	10.12±7.14

3.4 Results

In this section, we present the results of our experiments for two different tasks, quantified in terms of average mean squared error (MSE) and the average number of days. We present the results in terms of MSE as it is a commonly used metric in regression tasks and provides a quantitative measure of the deviation between the predicted values and the actual values. While we evaluate the performance of our approach specifically in the medical domain, it is important to note that our approach is domain-agnostic and can be applied to other domains. Hence, MSE provides a suitable evaluation metric for measuring the performance of our approach in a standardized manner.

We additionally present the results in terms of the number of days, so as to see the actual impact of our approach in these two tasks. Even though in terms of MSE the difference between our approach and other approaches is fairly minimal, the difference in days shows the actual impact of our approach. As in the medical domain differences of even 1 day are significant due to improving treatment options and outcomes.

To facilitate a comprehensive analysis of the results, we begin by showcasing the presence of missing data values in each dataset, as depicted in the first subsection. This serves as an essential context for understanding the subsequent findings and our interpretations.

3.4.1 Missing Data in the Datasets

In both the preterm birth dataset (PTB) and organ transplant dataset, missing values were observed, which is a common occurrence in medical data. These missing values arise due to various reasons. In the PTB task, the data is sourced from a survey, and some respondents did not provide answers to all the questions. On the other hand, the organ transplant dataset, collected through a national collaboration, includes samples with missing data. To handle the missing values in these datasets, we replaced them with a specific value. For both our datasets we hypothesized that the missing values would negatively impact our approach. Our experiments further confirmed our hypothesis. For the PTB task, when we finetune the models on a combination of PTB and NPTB data, we observed that higher missing values in the NPTB training dataset lowered the test performance of our model. The missing values in the NPTB dataset are shown in Table 3.2 and the organ transplant data in Table 3.3.

Table 3.2: Missing values in the NPTB train dataset across each fold for the Preterm Birth task

Fold1	Fold2	Fold3	Fold4	Fold5
183	178	184	179	178

3.4.2 Results in terms of average mean squared error (MSE)

We present the average mean squared error (MSE) over five cross-validations folds for each task. The results for the PTB task can be found in Table 3.4. Additionally,

Table 3.3: Missing values in the training dataset across each fold in the Organ Transplant task

Fold1	Fold2	Fold3	Fold4	Fold5
3732	3609	3554	3517	3364

we provide the standard deviation (SD) values across all baselines for each of the minifolds within this task. Similarly, the results for the Organ Transplant task are presented in Table 3.5.

Upon careful analysis, it is apparent that our proposed approach, QDTL, consistently outperforms all other baselines and ablations across both tasks. This empirical evidence substantiates Quality-Diversity as the superior choice among alternative search-based optimization techniques (specifically, our ablation techniques), addressing our third research question regarding the optimal selection of Quality-Diversity. Although the differences in MSE may appear small, they translate to a difference of multiple days in the predictive precision. These findings highlight the effectiveness of our approach in enhancing predictive accuracy, thus addressing our second research question.

On the PTB dataset, the greedy approach performs closest to our approach QDTL in terms of average MSE. The performance of the no-transfer learning model is significantly different from our proposed QDTL approach. Additionally, the no-transfer learning model yields lower performance than the basic finetuning approach. These findings underscore the importance of transfer learning for our task and further highlight the need for a more sophisticated approach than basic finetuning to achieve optimal results. While the CE-MCTS method shows some improvement compared to no-transfer learning, it fails to provide a significant improvement over the input model (i.e., finetuning model). Conversely, the results obtained using the greedy and random exploration approaches are superior, which corroborates our hypothesis that a more diverse exploration of the model space would help identify an optimal model.

Table 3.4: Average Mean Square Error (MSE) loss over five cross-validation folds of the PTB Task

Approach	Fold1	Fold2	Fold3	Fold4	Fold5	Average
No-Transfer Learning	0.101±0.020	0.145±0.036	0.115±0.008	0.137±0.065	0.205±0.066	0.141±0.039
Finetune	0.103±0.014	0.136±0.024	**0.114±0.013**	**0.103±0.004**	0.218±0.167	0.135±0.044
CE-MCTS	**0.095±0.013**	0.141±0.013	0.115±0.008	0.110±0.007	0.199±0.095	0.132±0.027
Random Walk	0.096±0.010	0.140±0.009	0.115±0.009	0.108±0.006	0.156±0.049	0.123±0.017
Greedy	**0.095±0.013**	0.137±0.003	0.115±0.007	0.110±0.007	0.151±0.032	0.122±0.012
GA	0.098±0.009	0.139±0.008	0.117±0.009	0.108±0.005	0.204±0.151	0.133±0.037
QDTL	0.100±0.010	**0.124±0.011**	0.118±0.001	0.107±0.005	**0.132±0.003**	**0.116±0.006**

Table 3.5: Average Mean Square Error (MSE) loss over five cross-validation folds of the Organ Transplant Task

Approach	Fold1	Fold2	Fold3	Fold4	Fold5	Average
No Transfer Learning	0.037	0.108	0.092	0.085	0.069	0.078±0.027
Finetune	0.037	0.110	0.100	0.097	0.070	0.083±0.030
QDTL	**0.036**	0.108	0.092	**0.080**	**0.065**	**0.076±0.027**

The finetuning approach outperforms QDTL on fold 3 and fold 4, while CE-MCTS and Combined Greedy approach outperform QDTL on fold 1. This can be attributed to the higher proportion of missing data in the training dataset (as shown in Table 3.2) for these folds, as compared to fold 2 and fold 5.

In the context of the Organ transplant task, we observe that our QDTL approach consistently outperforms both baselines on average. Notably, our approach exhibits significantly better performance than the finetune model, which serves as our base model for quality-diversity optimization. QDTL performs similarly to the no-transfer learning approach on fold2 and fold3, and nearly equally on fold1, due to the presence of more missing data in the training dataset in these three folds compared to the other two (as shown in table 3.3). Therefore, the performance of our approach is slightly reduced.

3.4.3 Results in Average Error in Days

In this section, we present the analysis of the average number of days by which the model is off for two different tasks: preterm birth (PTB) and organ transplant. The average number of days off is computed across all folds in each task.

Table 3.6 displays the average number of days off for each fold in the PTB task, while Table 3.7 represents the same for the organ transplant task.

To calculate the average number of days off for the PTB task, we considered each mini-fold within a fold and computed the average of the predicted days. For instance, in the case of fold1, we averaged the predicted days from fold1A, fold1B, fold1C, and fold1D. This procedure was repeated for each fold to determine the average number of days for the task. Subsequently, we compared the average number of days to the actual number of days by calculating the absolute difference. We then divided this absolute difference by the number of test samples in the particular fold. It is worth noting that the test dataset contains 9 samples for fold1, fold2, and fold3, while for fold4 and fold5, there are 8 samples each.

For the organ transplant task, since there are no mini-folds, we directly computed the absolute difference between the actual and predicted number of days and divided this absolute difference by the number of test samples. Which is 26 in this case.

Upon analyzing the PTB task results, it is evident that the Quality-Diversity Transfer Learning (QDTL) approach performs exceptionally well in fold2 and fold5, with an approximate difference of 2 days compared to other approaches. A difference of 2 days is significant for our problem. In most cases, medical interventions for PTB need to be given within a week of likely birth. Although Greedy and Finetuning perform better on fold1, fold3, and fold4 as compared to QDTL, on average, there is only a one-day difference between QDTL and the other approaches. Across these three folds, it is also noticeable that the NPTB training dataset (as shown in Table 3.2) contains more missing data.

Table 3.6: Average number of days by which the model is off over five cross-validation folds of PTB Task

Approach	Fold1	Fold2	Fold3	Fold4	Fold5
No-Transfer Learning	9.181	3.061	17.289	4.908	18.8
Finetune	9.274	3.054	**16.854**	**4.902**	21.144
CE-MCTS	9.197	3.128	17.192	5.273	21.949
Random Walk	9.242	3.126	17.243	5.259	20.491
Greedy	**9.181**	3.060	17.289	5.296	19.990
GA	9.249	3.156	17.230	5.222	21.094
QDTL	9.380	**2.792**	17.886	5.179	**18.382**

In the organ transplant task, the performance of the QDTL approach outperforms the other two baselines in all folds except for fold2. It is important to note that the numbers reported for this task are relatively high due to the nature of the data, where we are predicting over years instead of within a single year.

In this task, each test fold consists of 26 samples, which were used for evaluating the performance of the different approaches. The QDTL approach demonstrates superior performance compared to the other baselines, indicating its effectiveness in predicting the survival days for organ transplant patients across most folds. However, it is observed that in fold2, the QDTL approach did not perform as well as the no-transfer learning approach, but the difference is relatively small.

In the majority of folds, finetuning performs worse than no-transfer. This indicates that for naive backpropagation this problem represents an example of negative transfer. Despite this, and the fact that we use the finetuning model as input, our approach outperforms no-transfer learning in all but one fold. This indicates that QDTL can negate the effects of negative transfer. It also suggests that in cases with negative transfer it might be wise to use the no-transfer learning model as input for QDTL rather than finetune.

Table 3.7: Average number of days by which the model is off over five cross-validation folds of Organ Transplant Task

Approach	Fold1	Fold2	Fold3	Fold4	Fold5
No-transfer Learning	169.660	**213.301**	216.500	168.357	266.028
Finetune	167.057	214.798	226.761	185.142	275.055
Combined QD	**162.697**	213.485	**216.080**	**163.765**	**246.677**

3.4.4 Statistical Tests

Statistical tests provide insights into the performance of a model. The p-values obtained from the tests indicate the statistical significance of any observed differences in performance. In this analysis, we examine the p-values for paired t-tests conducted between our approach QDTL, baselines and ablations across the five-folds, allowing us to assess the significance of the differences and gain a comprehensive understanding of the comparative performance. Table 3.8 represents the p-values of paired t-tests comparing the MSE values across each of the folds between QDTL and our baselines and ablations. From the results, we can see that there are significant performance differences between QDTL and two of the ablations in fold 1 (e.g., Random Walk in Fold1, Greedy in Fold1), while in other folds and for other approaches, the performance differences are not significant. The p-values for both No-Transfer Learning and Finetune baselines were relatively high in all folds, indicating no significant difference in performance compared to QDTL. However, we know there were large differences in terms of error. Thus we interpret these results to indicate that there were similar predictions being made, but that QDTL was more precise.

Table 3.8: Paired t-test p-values comparing the MSE observed across 5 cross-validation folds between QDTL and the different baseline approaches. Bold values are significant

Approach	Fold1	Fold2	Fold3	Fold4	Fold5
No-Transfer Learning	0.875	0.413	0.685	0.426	0.105
Finetune	0.333	0.492	0.944	0.398	0.759
CE-MCTS	0.055	0.084	0.556	0.655	0.253
Random Walk	**0.005**	0.202	0.619	0.776	0.392
Greedy	**0.039**	0.124	0.574	0.615	0.306
GA	0.069	0.218	0.970	0.826	0.406

Chapter 4

Conclusion

In this concluding chapter, we summarize the key findings and contributions of this thesis. Section 4.1 discusses the implications of our work, highlighting the potential impact and relevance in the respective domains. We also address the limitations encountered during the research process and discuss potential avenues for future work in Section 4.2. Lastly, Section 4.3 provides closing thoughts, reflecting on the overall significance and implications of this work.

4.1 Implications

This research introduced an innovative approach called Quality-Diversity Transfer Learning (QDTL) to address the limitation of data in medical domain prediction tasks. The effectiveness of the QDTL approach is evaluated by applying it to two distinct medical domain tasks. The findings demonstrate that our approach meets or exceeds the performance of baseline models and ablations in terms of performance, particularly with low missing values.

Our findings regarding our first research question, which investigates whether our approach can outperform standard logistic regression, are presented in Table 3.1. The findings clearly indicate that the finetuning approach yields better results than logistic regression. Our QDTL approach, as depicted in Table 3.4 and Table 3.5, exhibits even superior performance compared to the finetuning approach. Furthermore, the results in Table 3.6 and Table 3.7 indicate that our QDTL approach surpasses other transfer learning techniques for predictive care in the medical domain, addressing our second research question. Finally results in Table 3.4 and Table 3.5 show that quality diversity optimization performs better than other search-based techniques. Hence addressing our third research question. In summary, our findings provide compelling evidence supporting the effectiveness of our approach over standard logistic regression, as well as other search-based techniques, in addressing predictive care challenges in the medical domain.

4.2 Limitations and Future Work

The results demonstrate the efficacy of our approach in this domain, suggesting the potential for it to address low-data problems in other domains. However, further exploration is needed. In addition, the performance of our approach is currently dependent on the finetuned model used as input, and the finetuning process requires careful consideration of the dataset, model architecture, and other features of the source and target domains. This dependence on the finetuning process creates challenges in terms of choosing various hyperparameters, which we plan to investigate in future work. Overall, our study contributes to the ongoing research on addressing limited data challenges in medical prediction models and paves the way for exploring the application of QDTL in other low-data domains.

4.3 Closing Thoughts

In this thesis, we present a novel approach called Quality-Diversity-Transfer Learning (QDTL) to solve low-data problems. This approach relies on a combination of transfer learning, architecture search and an evolutionary approach. We evaluate the effectiveness of our approach by comparing it with standard baselines and ablations on two medical prediction tasks. Our results indicate that QDTL outperforms the baselines and ablations in terms of efficiency, resulting in higher-quality models more closely able to predict desired output features. This demonstrates the potential of our approach to improve model performance on various low-data tasks.

Bibliography

[1] I. Sarker, *Deep learning: A comprehensive overview on techniques, taxonomy, applications and research directions. sn comput. sci. 2, 420 (2021).*

[2] P. Villalobos, J. Sevilla, L. Heim, T. Besiroglu, M. Hobbhahn, and A. Ho, "Will we run out of data? an analysis of the limits of scaling datasets in machine learning," *arXiv preprint arXiv:2211.04325*, 2022.

[3] S. Shamshirband, M. Fathi, A. Dehzangi, A. T. Chronopoulos, and H. Alinejad-Rokny, "A review on deep learning approaches in healthcare systems: Taxonomies, challenges, and open issues," *Journal of Biomedical Informatics*, vol. 113, p. 103 627, 2021.

[4] I. Bica and M. van der Schaar, *Transfer learning on heterogeneous feature spaces for treatment effects estimation*, 2022. arXiv: 2210.06183 `[cs.LG]`.

[5] M. Yoon, J. Palowitch, D. Zelle, Z. Hu, R. Salakhutdinov, and B. Perozzi, *Zero-shot transfer learning within a heterogeneous graph via knowledge transfer networks*, 2022. arXiv: 2203.02018 `[cs.LG]`.

[6] K. Weiss, T. M. Khoshgoftaar, and D. Wang, "A survey of transfer learning," *Journal of Big data*, vol. 3, no. 1, pp. 1–40, 2016.

[7] M. Guzdial and M. O. Riedl, "Combinets: Creativity via recombination of neural networks," *International Conference on Computational Creativity*, 2019.

[8] A. Mahajan and M. Guzdial, "Modeling individual humans via a secondary task transfer learning method," in *Federated and Transfer Learning*, Springer, 2022, pp. 259–281.

[9] M. Singamsetti, A. Mahajan, and M. Guzdial, "Conceptual expansion neural architecture search (cenas)," *International Conference on Computational Creativity*, 2021.

[10] V. Nair and G. E. Hinton, "Rectified linear units improve restricted boltzmann machines," in *Proceedings of the 27th international conference on machine learning (ICML-10)*, 2010, pp. 807–814.

[11] Y. Yao and Z. Huang, "Bi-directional lstm recurrent neural network for chinese word segmentation," in *Neural Information Processing: 23rd International Conference, ICONIP 2016, Kyoto, Japan, October 16–21, 2016, Proceedings, Part IV 23*, Springer, 2016, pp. 345–353.

[12] C. Tan, F. Sun, T. Kong, W. Zhang, C. Yang, and C. Liu, “A survey on deep transfer learning: 27th international conference on artificial neural networks, rhodes, greece, october 4–7, 2018, proceedings, part iii,” in Oct. 2018, pp. 270–279, ISBN: 978-3-030-01423-0. DOI: 10.1007/978-3-030-01424-7_27.

[13] Y. Guo, H. Shi, A. Kumar, K. Grauman, T. Rosing, and R. Feris, “Spottune: Transfer learning through adaptive fine-tuning,” in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4805–4814.

[14] M. A. Boden, “Creativity and artificial intelligence,” *Artificial intelligence*, vol. 103, no. 1-2, pp. 347–356, 1998.

[15] M. Tomassini, “Evolutionary algorithms,” in *Towards evolvable hardware: the evolutionary engineering approach*, Springer, 2005, pp. 19–47.

[16] J.-B. Mouret and G. Maguire, “Quality diversity for multi-task optimization,” in *Proceedings of the 2020 Genetic and Evolutionary Computation Conference*, 2020, pp. 121–129.

[17] L. Keller, D. Tanneberg, S. Stark, and J. Peters, “Model-based quality-diversity search for efficient robot learning,” in *2020 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, IEEE, 2020, pp. 9675–9680.

[18] D. Gravina, A. Khalifa, A. Liapis, J. Togelius, and G. N. Yannakakis, “Procedural content generation through quality diversity,” in *2019 IEEE Conference on Games (CoG)*, IEEE, 2019, pp. 1–8.

[19] D. Floreano, P. Dürr, and C. Mattiussi, “Neuroevolution: From architectures to learning,” *Evolutionary intelligence*, vol. 1, pp. 47–62, 2008.

[20] G. F. Miller, P. M. Todd, and S. U. Hegde, “Designing neural networks using genetic algorithms.,” in *ICGA*, vol. 89, 1989, pp. 379–384.

[21] X. Liu, J. Zhao, J. Li, B. Cao, and Z. Lv, “Federated neural architecture search for medical data security,” *IEEE transactions on industrial informatics*, vol. 18, no. 8, pp. 5628–5636, 2022.

[22] H. Guo, X. Zhuang, P. Chen, N. Alajlan, and T. Rabczuk, “Stochastic deep collocation method based on neural architecture search and transfer learning for heterogeneous porous media,” *Engineering with Computers*, vol. 38, no. 6, pp. 5173–5198, 2022.

[23] R. Romero, S. K. Dey, and S. J. Fisher, “Preterm labor: One syndrome, many causes,” *Science*, vol. 345, no. 6198, pp. 760–765, 2014.

[24] H. H. Chang *et al.*, “Preventing preterm births: Analysis of trends and potential reductions with interventions in 39 countries with very high human development index,” *The Lancet*, vol. 381, no. 9862, pp. 223–234, 2013.

[25] R. M. Patel, “Short-and long-term outcomes for extremely preterm infants,” *American journal of perinatology*, vol. 33, no. 03, pp. 318–328, 2016.

[26] Y. J. Heng *et al.*, “Maternal whole blood gene expression at 18 and 28 weeks of gestation associated with spontaneous preterm birth in asymptomatic women,” *PloS one*, vol. 11, no. 6, e0155191, 2016.

[27] M. N. Moufarrej *et al.*, "Early prediction of preeclampsia in pregnancy with cell-free rna," *Nature*, vol. 602, no. 7898, pp. 689–694, 2022.

[28] K. A. Scott, B. D. Chambers, R. J. Baer, K. K. Ryckman, M. R. McLemore, and L. L. Jelliffe-Pawlowski, "Preterm birth and nativity among black women with gestational diabetes in california, 2013–2017: A population-based retrospective cohort study," *BMC pregnancy and childbirth*, vol. 20, no. 1, pp. 1–14, 2020.

[29] M. Bhat, M. Rabindranath, B. S. Chara, and D. A. Simonetto, "Artificial intelligence, machine learning, and deep learning in liver transplantation," *Journal of hepatology*, vol. 78, no. 6, pp. 1216–1233, 2023.

[30] I. Christiaens, K. Hegadoren, and D. M. Olson, "Adverse childhood experiences are associated with spontaneous preterm birth: A case–control study," *BMC medicine*, vol. 13, pp. 1–9, 2015.

[31] M. Maqsood *et al.*, "Transfer learning assisted classification and detection of alzheimer's disease stages using 3d mri scans," *Sensors*, vol. 19, no. 11, p. 2645, 2019.

[32] H.-C. Shin *et al.*, "Deep convolutional neural networks for computer-aided detection: Cnn architectures, dataset characteristics and transfer learning," *IEEE transactions on medical imaging*, vol. 35, no. 5, pp. 1285–1298, 2016.

[33] M. Jovanovic, S. Radovanovic, M. Vukicevic, S. Van Poucke, and B. Delibasic, "Building interpretable predictive models for pediatric hospital readmission using tree-lasso logistic regression," *Artificial intelligence in medicine*, vol. 72, pp. 12–21, 2016.