ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 2 - 2026

A migration learning-based approach to cross-domain model building for biological big data analytics

Xinxin Gan[1] *and Linhui Wang*[2,*]

[1] School of Materials Science and Engineering, University of Shanghai for Science and Technology, Shanghai, 200082, China
[2] Department of Urology, Changhai Hospital, Naval Medical University, Shanghai, 200082, China

SUMMARY: *A cross-domain model based on migration learning can solve the problems of missing annotated data and large distributional differences between domains, these are prevalent in biological data fields, like genomics, proteomics, and medical imaging. In this research paper, a novel deep transfer learning model founded on multi - source domain integration (MUCT) is put forward, building upon conventional cross - domain transfer learning approaches. Firstly, an end-to-end training mechanism is established based on deep neural networks, secondly, high-confidence target samples collected through consistency filters are trained as a way to create target domain supervisory information, and finally, the outcomes of the classification achieved among multiple source domains and the target domain are combined by means of the relative majority voting approach to enhance the model's resilience. This approach demonstrates an excellent identification outcome for medical entities within Chinese electronic medical records, with a strict F1 value of 85.4% on the CCKS 2018 review dataset. Typical case study results validate that the migration method can effectively recognize entities such as personal information, disease symptoms, diagnosis and treatment, and drug use in patient question texts by utilizing only a small amount of annotated corpus, realizing the full utilization of existing data resources. This study provides an efficient knowledge migration paradigm for biological big data analysis, which is expected to promote the in-depth development of precision medicine and systems biology research.*

KEYWORDS: *migration learning; cross-domain modeling; biological data; MUCT; deep neural network*

1 Introduction

In the 21st century, with the continuous development of life sciences, bioinformatics has entered the post - genomic era, characterized by the comprehensive sequencing across the entire genome. At present, its research undertakings primarily focus on large - scale genomic analysis, proteomic analysis, as well as the analysis and identification of the causative genes of complex diseases[1-3]. Most of today's biomedical researchers use a DNA sequencing method called high-throughput sequencing, this enables the concurrent sequencing of hundreds of thousands, or even millions, of DNA molecules. [4]. At the same time, the amount of genetic data generated has increased geometrically due to the dramatic increase in the number of biogenetic experiments [5, 6]. The 21st century has undoubtedly become the era of biological big

*ganxxhee@163.com
https://doi.org/10.65102/is2026610
Published: 30 April, 2026

data.Within this research environment, large - scale data analysis is emerging as the next significant area in the realm of bioinformatics. It combines data archiving, data dissemination, data examination, and data quality management to foster the growth of the large - data application sector and create entirely novel prospects [7-9].

However, biological big data analysis is often faced with problems such as difficult data acquisition, differentiated data distribution, high annotation cost, etc. Analyzing large-scale biological data using conventional machine learning models poses challenges in discerning data characteristics when the available data is scarce, and the generalization ability is insufficient, resulting in incomplete capture of biological laws [10-13]. In contrast, transfer learning, a technique within the realm of machine learning, boosts the efficiency and effectiveness of a target task. It does so by leveraging the knowledge acquired from one task and applying it to a related task.The fundamental concept revolves around the utilization of the source task. the existing data and models to reduce the amount of labeled data and training time required for the target task, which can effectively solve the problems of insufficient data and high labeling cost [14]. Migration learning has a large potential for application in biological big data analysis, which provides support for biological research.

Rogers et al. demonstrated through case comparisons that transfer learning effectively integrates multi-source data of biochemical processes, improves prediction accuracy, and reveals the association between process mechanisms and model structure, enhancing model interpretability [15]. Theodoris et al. introduced a pre-training model based on migration learning, which can be fine-tuned with limited samples by pre-training with massive single-cell transcriptome data, significantly improving the prediction accuracy of gene networks, and has been successful in identifying potential therapeutic targets for cardiomyopathies [16]. Ashayeri et al. reused pre-trained models through migration learning, which significantly improved the prediction accuracy and efficiency in biological data tasks such as mutation detection, expression analysis, and identification of genetic syndromes, and is a key technique to optimize artificial intelligence applications for genetic research [17]. Zhang and colleagues put forward a transfer learning approach for small - scale genotype - phenotype data. This was achieved by adjusting the weights of the pre - trained model and carrying out association testing, which effectively improved the prediction accuracy and candidate gene identification in nicotine dependence cases [18]. Muneeb et al. With the help of migration learning, this method improved the classification accuracy by 2%-14.2%, which provides new ideas for endangered species research and precision medicine [19]. Giorgi and Bader used migration learning to migrate models trained on a large-scale noisy standard corpus to a small-scale high-quality gold-standard corpus, resulting in an average reduction in biomedical named-entity recognition error rates of about 11%, especially for label-sparse datasets [20]. Kensert et al. applied a deep convolutional neural network (CNN) model for migration learning to classify cellular morphology images, which significantly reduces the dependence on large amounts of labeling data and demonstrates its efficiency in high-throughput phenotyping [21]. Pickering et al. also introduced CNN for transfer learning for analyzing African forest elephant acoustic data, which classifies calls and identifies subtypes with high accuracy, effectively captures biologically relevant acoustic changes such as age and behavior, and outperforms traditional feature engineering [22]. Zheng et al. develop a novel transfer learning algorithm for motor imagery EEG classification, which achieves a test accuracy of 85.7% across subjects and across experimental scenarios by extracting and updating inter-session shared features, significantly outperforming traditional machine learning methods [23].

In addition, a number of researchers have investigated cross-domain model construction methods for transfer learning in the biological domain. Bird et al. verified the feasibility of unsupervised transfer learning between EMG and EEG signals by optimizing the neural

network to achieve cross-domain knowledge transfer, and EMG to EEG migration increased the classification accuracy to 93.82%, indicating that this strategy can effectively reduce the need for complex model construction [24]. Gu et al. proposed two approaches, progressive migration and adversarial learning, to address the cross-domain recognition challenges of skin disease images, which were verified to be effective in improving the generalization ability of the model and alleviating the domain bias problem through testing on heterogeneous clinical datasets [25]. Guo et al. explored deep migration learning between different biological image domains with significant species differences, and proposed a multi-stage cross-domain migration method with the introduction of intermediate domains in response to the poor migration results caused by excessive differences between the source and target domains [26]. Yan et al. proposed a deep transfer learning framework for cross-species diagnosis of plant diseases, which effectively improves the effect of knowledge transfer between weakly related domains by generating mixed-domain images and introducing a subdomain alignment mechanism, and outperforms the existing methods in fine-grained information capture [27]. Maswanganyi et al. formulated an enhanced multi-source streaming feature transfer learning framework to solve the negative migration problem in the cross-object and cross-discipline classification of multi-class EEG signals. Experiments show that the method can effectively reduce the effect of negative migration with a maximum classification accuracy of 98% [28]. Li et al. created a deep neuro-fuzzy system called “Fuzzy-ViT”, which efficiently transforms generic features extracted by a pre-trained visual transformer into medical features through a fuzzy-attention cross-domain module, and realizes efficient cross-domain transfer learning on medical image benchmarks [29].

This paper systematically investigates a cross-domain model construction method based on migration learning, aiming to solve the core bottleneck in biological data analysis. The study combines distributional adaptive methods and feature selection methods to propose a deep migration learning model based on multi-source domain integration (MUCT). The Resume dataset is selected for pre-training and initializing the network model, after which the electronic medical record dataset is used to train the model to further improve the learning ability of the whole network model. Finally, comparative experiments are conducted on the CCKS 2018 review dataset, and the effectiveness of the construction method is verified with typical case studies.

2 Method

Migration learning can effectively alleviate the problem of data scarcity by migrating knowledge from the source domain to the target domain. In bioinformatics, cross-domain migration learning has important applications: for example, using annotation data of model organisms (e.g., mice) to assist human disease research, utilizing data from different sequencing platforms to perform complementary analysis, and migrating knowledge of normal tissues to disease tissue analysis. In this chapter, traditional cross-domain migration learning methods are investigated, and a deep migration learning model (MUCT) based on multi-source domain integration is proposed in combination with data distribution adaptation and feature selection methods, as a way to realize accurate analysis of biological big data.

2.1 Traditional cross-domain transfer learning methods

2.1.1 Data Distribution Adaptive Methods

The fundamental concept of data distribution adaptation is that given the disparity in the

probability of data distribution between the source and target domains, a specific transformation is employed to directly reduce the gap between the two distributions. These methods can be classified into marginal distribution adaptation, conditional distribution adaptation, and joint distribution adaptation.

(1) Marginal distribution adaptation

Reduce the difference between the edge probability distributions of the source and target domains. Migrating component analysis (TCA) is the earliest adaptive method for edge distribution, TCA assumes that there exists a feature mapping θ, which makes the distribution of the data after the mapping approximately the same, and the method employs the maximum mean discrepancy distance to determine the disparity in the means of the source domain and the target domain subsequent to the mapping. By minimizing the MMD distance while preserving the data characteristics of the source and target domains respectively, it brings the data distribution closer to that of the target domain. When dealing with two sets of data having distinct probability distributions, TCA processing results in a closer alignment of these probability distributions.

(2) Conditional Probability Distribution Adaptation

This approach aims to minimize the disparity between the conditional probability distributions of the source and target domains. It is primarily employed when the data in the source domain possess category labels, while the data in the target domain lack such labels. Conditional distribution adaptation mainly uses the distance between $P\left(y_s \middle| x_s\right)$ and $P\left(y_t \middle| x_t\right)$ to approximate the difference between two domains:

$$dis\tan ce\left(Domain_{source}, Domain_{target}\right) \approx +\left\|P\left(y_s \middle| x_s\right) - P\left(y_t \middle| x_t\right)\right\| \tag{1}$$

(3) Joint Distribution Adaptation

Reducing the distance of the joint distribution of the source and target domains is approximated by the distance between $P(x_s)$ and $P(x_t)$, as well as the distance between $P\left(y_s \middle| x_s\right)$ and $P\left(y_t \middle| x_t\right)$ to approximate the difference between the two fields, i.e., the distance between:

$$\begin{aligned} dis\tan ce\left(Domain_{source}, Domain_{target}\right) \approx & \left\|P\left(x_s\right) - P\left(x_t\right)\right\| \\ & + \left\|P\left(y_s \middle| x_s\right) - P\left(y_t \middle| x_t\right)\right\| \end{aligned} \tag{2}$$

The most representative of these is the JDA (Joint Distribution Adaptation) method, where the goal of JDA is to find a transformation A such that the transformed distances of $P\left(A^T x_s\right)$ and $P\left(A^T x_t\right)$ can be as close as possible, while the distances of $P\left(y_s \middle| A^T x_s\right)$ and $P\left(y_t \middle| A^T x_t\right)$ are also smaller.

2.1.2 Feature selection methods

The feature selection approach presupposes that both the source and target domains possess a segment of shared features where the data distributions of the source and target domains are in agreement. The SCL (Structural Correspondence Learning) approach is among the typical algorithms of the feature selection approach, which can transform some features unique to a

space to the axis features shared by all domains by mapping, and then use machine learning algorithms to make classification predictions on the axis features. The underlying principle of the SCL method is depicted in Figure 1. In the center of the figure, the horizontal axis represents the shared characteristic axis of two domains: MEDLINE and WSJ. The characteristics in the upper - half of the figure are exclusive to MEDLINE, while those in the lower - half are specific to WSJ.The vertical axis in the middle divides these characteristics into two categories. The characteristics in the left - hand half of the figure occur frequently in the positive category. On the contrary, the characteristics in the right - hand half occur frequently in the negative category.By constructing a classifier on the shared characteristic axis, the classifier can leverage the features from the source domain to make classification determinations regarding the target domain data.

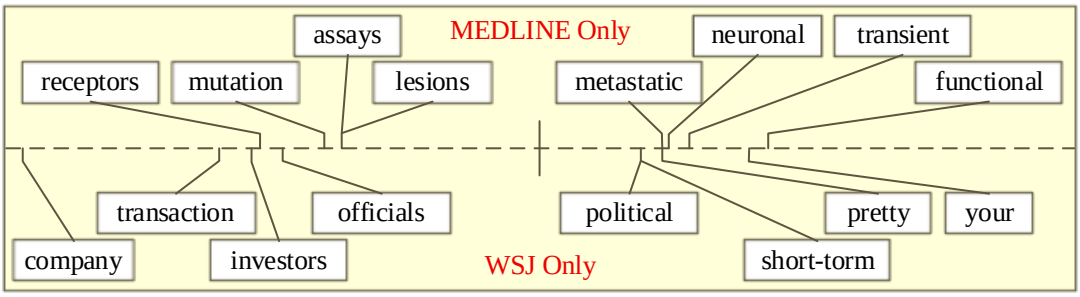


Figure 1: Principle of SCL method

2.2 Deep migration learning model based on multi-source domain integration

2.2.1 Model Architecture

In this paper, building upon the conventional cross - domain migration learning approach, specifically data distribution adaptation, an adaptive layer is constructed within deep learning. This layer is designed to achieve the self - adjustment of data across the source and target domains. The adaptive layer computes the migration loss that occurs when moving from the source domain to the target domain. By minimizing this loss, the data distributions of the source and target domains become more similar. As a result, the network gains the capacity to automatically adapt to diverse domains. And based on the traditional cross-domain migration learning method, the feature selection method, the feature selection layer is proposed to establish a screening mechanism for the source domain features. The deep migration learning model (MUCT) architecture based on multi-source domain integration constructed Figure 2 presents what is demonstrated in this paper.

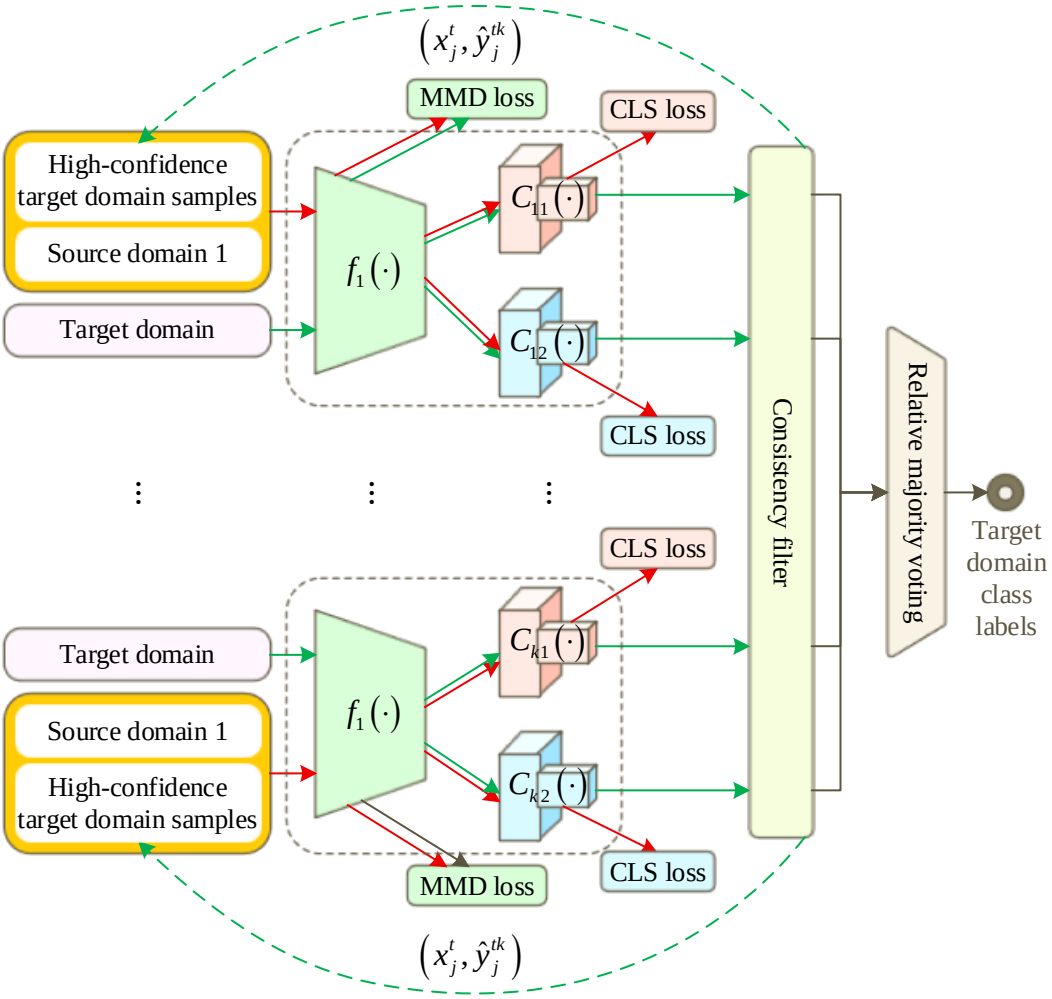


Figure 2: MUCT model architecture

2.2.2 Model construction methodology

In this segment, we'll commence by delineating the issue explored in this chapter. Subsequently, we'll detail the advanced domain distribution alignment approach, high-confidence sample screening approach, and model learning and fusion strategy involved in the model, respectively.

(1) Notation definition and problem description

This chapter focuses on a multi-source domain unsupervised transfer learning task, where given K source domains $\mathcal{D}_s = \{\mathcal{D}_{sk}\}_{k=1}^{K}$, notate the k th source domain as $\mathcal{D}_{sk} = \left\{\left(x_i^{sk}, y_i^{sk}\right)\right\}_{i=1}^{n_{sk}}$, which contains n_{sk} samples of the source domains with labeled information, and notate the target domain as $\mathcal{D}_t = \left\{x_j^t\right\}_{j=1}^{m}$, which contains m unlabeled target domain samples. We assume that $\mathcal{X}_{sk} = \mathcal{X}_t$ and $\mathcal{Y}_{sk} = \mathcal{Y}_t$] but that the data distributions are different between the source domains, and that $\mathcal{D}_{sk}$ and $\mathcal{D}_t$ are both from different data distributions, denoted as $P(X_{sk}) \neq P(X_t)$. The goal of this chapter is to build the final classification network model $H(\cdot)$ by means of a multi-model combination that minimizes the difference between the distributions of the different source domains and the target domains while at the same time using the source domain supervisory information to train the classifiers

so as to minimize the discrepancy between the model's prediction on the target domain and its true value $R_t = \mathbb{E}_{(x,y)\in D_t}\left[H(x) \neq y\right]$.

(2) Deep domain distribution alignment method

Since the neural network model will gradually focus from general public features over features with specific tasks or domains as the number of network layers gradually deepens when extracting sample features, deep domain adaptation usually uses a public feature extractor f to share the model parameters of the shallow network. Drawing on the concept that the maximum limit of the classifier's error within the target domain primarily stems from the disparity between the model's error in the source domain and the domain's distribution, the general optimization objective of deep domain adaptation can be formalized as:

$$\min \frac{1}{n_s}\sum_{i=1}^{n_s}\mathcal{L}_{ce}\left(f\left(x_i^s\right), y_i^s\right) + \lambda \hat{d}\left(P_s, P_t\right) \tag{3}$$

where $\mathcal{L}_{ce}$ denotes the cross-entropy loss, which is used to compute the classification loss of the model on the source domain, $\hat{d}$ is used to measure the distributional difference between the source and target domains, and λ is the trade-off parameter between the domain adaptation loss and the classification loss.

For the measure of distributional difference, we adopt MMD.As a nonparametric measure, MMD does not require any a priori assumptions on the distributions and can compare the distance between two distributions without knowing their forms. Specifically, MMD measures the difference in distributions between the source and target domains via mean embedding in RKHS in kernel space. Formally, the MMD distance approach for samples from distribution P_s and distribution P_t is as follows:

$$d_{mmd}\left(P_s, P_t\right) \triangleq \left\| \mathbb{E}_{x^s \sim P_s}\left[\varphi\left(x^s\right)\right] - \mathbb{E}_{x^t \sim P_t}\left[\varphi_t\left(x^t\right)\right] \right\|_{\mathcal{H}}^2 \tag{4}$$

where $\varphi(\cdot)$ is the mapping function that, by using the feature kernel k, $k\left(x^s, x^t\right) = \left\langle \varphi\left(x^s\right), \varphi\left(x^t\right) \right\rangle$, which maps source domain samples x^s and target domain samples x^t into the regenerative kernel Hilbert space $\mathcal{H}$. $P_s = P_t$ if and only if $d_{mmd}\left(P_s, P_t\right) = 0$. The empirical estimate of MMD $\hat{d}_{mmd}\left(P_s, P_t\right)$ can be further decomposed as:

$$\begin{aligned} \hat{d}_{mmd}\left(P_s, P_t\right) &= \left\| \frac{1}{n}\sum_{x_i \in D_s}\varphi\left(x_i^s\right) - \frac{1}{m}\sum_{x_j \in D_t}\varphi\left(x_j^t\right) \right\|_{\mathcal{H}}^2 \\ &= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{l=1}^{n}k\left(x_i^s, x_l^s\right) + \frac{1}{m^2}\sum_{j=1}^{m}\sum_{l=1}^{m}k\left(x_j^t, x_l^t\right) \\ &\quad - \frac{2}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m}k\left(x_i^s, x_j^t\right) \end{aligned} \tag{5}$$

When dealing with the multi - source transfer learning issue, the construction of an appropriate common latent feature space presents a difficult challenge. This is because the data distributions across different domains vary. However, When the quantity of source domains

goes up, constructing a common shallow feature space among multiple domains leads to the loss of useful information in a single domain for the categorization of the target realm, which reduces the separability of features. Therefore, to enhance the ability of the feature extractor to characterize features, we employ K sub-networks to construct a common latent feature space for each source and target domain combination. Thus, for the domain difference loss under multiple source domains can be expressed as:

$$\hat{d}_{ms} = \sum_{k=1}^{K} \hat{d}_{mmd}\left(P_{s_k}, P_t\right) \tag{6}$$

(3) High-confidence sample screening approach

We propose a consistency filter to screen high-confidence target domain samples and assign pseudo-labels to them by predicting consistency under multiple viewpoints, so as to enrich the diversity of training samples and facilitate the learning process of target information. We consider that when the classification results from different viewpoints are the same, the target domain sample is considered to have high confidence, and the high-confidence target domain sample with pseudo-class labeling is added to the training set for the next round of model training. Therefore, the way to construct the classification from different classification perspectives given the high-confidence pseudo-class labeling by two heterogeneous classifiers C_{k1} and C_{k2} is:

$$\begin{aligned} \hat{y}_j^k &= \arg\max_c \, p_{k1}\left(y_j^k = c \middle| x_j^t\right) \\ &= \arg\max_c \, p_{k2}\left(y_j^k = c \middle| x_j^t\right) \end{aligned} \tag{7}$$

where $p_1\left(y_j^k \middle| x_j^t\right)$ and $p_2\left(y_j^k \middle| x_j^t\right)$ are the classifier C_{k1} and C_{k2} for the output probabilities of the extracted target domain features $f\left(x_j^t\right)$.

After filtering the target domain samples with high confidence, we add this sample $\left\{\left(x_j^t, \hat{y}_j^k\right)\right\}$ to the source domain to construct a new training set and proceed to the next round of model training. In order to prevent the classifier from not converging and causing some wrong pseudo-class labels to be added to the model training, we use the Warm-up operation, i.e., we run the consistency filter after a certain number of iterations for model training. The classification loss obtained based on the z th classifier under the source domain k is denoted as:

$$\mathcal{L}_{kz} = \frac{1}{n^{train}} \sum_{i=1}^{n^{train}} \mathcal{L}_{ce}\left(C_{kz}\left(f_k\left(x_i^{train}\right)\right), y_i^{train}\right) \tag{8}$$

The final total classification loss under K source domains is calculated as follows:

$$\mathcal{L}_{cl} = \sum_{k=1}^{K} \sum_{z=1}^{2} \mathcal{L}_{kz} \tag{9}$$

(4) Model learning and fusion strategy

In summary, a total of K feature extractors f with $2K$ classifiers c are included in

the model. The overall loss of the model is composed of domain loss and classification loss. The domain loss is gauged using the Maximum Mean Discrepancy (MMD), and the classification loss is computed via cross - entropy. Therefore, we express the loss function of the deep domain adaptive based multi-source domain integrated transfer learning model as a linear sum of the above two objectives:

$$\mathcal{L}_{total} = \mathcal{L}_{cl} + \lambda \hat{d}_{ms} \tag{10}$$

By leveraging the above equation for end - to - end training of the model, the model can extract the shared features of the source and target domains. Simultaneously, it guarantees the performance of these features on the final classifier. This effectively addresses the under - adaptation issue present in traditional domain adaptive approaches.Specifically, optimizing the domain loss empowers the feature extractor to capture the loss of domain invariance between the source and target domains. On the one hand, optimizing the classification loss minimizes the classification error on the source domain. On the other hand, by incorporating high - confidence target domain samples, it enhances the model's learning capacity for the target domain, thus resolving the underfitting problem of the model on the target domain.When it comes to generating the final classification results, we adopt the method of relative majority voting as described in equation (11) to obtain the ultimate classification outcomes:

$$H\left(x\right) = l_{\arg\max_{c} \sum_{k=1}^{K} \sum_{z=1}^{2} C_{kz}\left(f_k\left(x_j^t\right)\right)} \tag{11}$$

where l is the final class label obtained by relative majority voting for the test sample x_j^t.

3 Results and Discussion

To validate the efficacy of the cross - domain model building approach founded on transfer learning that is developed in this paper, the study designs experiments to conduct medical named entity recognition research on the model on the one hand, Conversely, the practical implementation outcome of the model is evaluated via representative cases.

3.1 Performance test experiment analysis

3.1.1 Experimental data set

The experiments use the Resume dataset and the CCKS 2018 review dataset. Among them, the Resume dataset is a non-medical dataset, which is used for the pre-training work of the MUCT model.The CCKS 2018 dataset is the dataset used for conducting the medical named entity recognition study in this chapter. These two datasets are described specifically below.

The Resume dataset consists of resumes of executives of Chinese stock market listed companies from Sina Finance. This dataset contains 1033 CV abstracts and eight categories of named entities, which are name of person (NAME), country (CONT), title (TITLE), education (EDU), organization (ORG), race (RACE), location (LOC), and profession (PRO), and the statistical results of each entity category are shown in Table 1.

Table 1: Resume data set entity statistics

Tags	NAME	CONT	TITLE	EDU	ORG	RACE	LOC	PRO
Quantity	1136	352	7652	1066	5641	132	86	365

The CCKS 2018 dataset is a dataset for the named entity recognition assessment task for Chinese EHRs, which mainly consists of inpatient medical records in EHRs, including inpatient case home page, admission records, course records, and pathology information. This dataset contains five entity types: anatomical site, symptom description, independent symptom, medication and surgery, and the entity statistics are shown in Table 2. Within this context, the training dataset consists of 500 present medical history records, while the test dataset comprises 300 present medical history records.

Table 2: CCKS 2018 data set entity statistics

Dataset	Document	Anatomy	Symptom	Independent	Drugs	Operation	All
Train	500	7963	2241	3036	1066	1145	15951
Test	300	6524	944	1258	822	736	10584

3.1.2 Evaluation indicators

The system's performance is assessed in the experiments by means of precision (P), recall (R), and the F1 score. In order to comprehensively reflect the overall performance of the model in predicting entity types, this paper follows the evaluation metrics of the CCKS 2018 evaluation task and analyzes them in terms of strict metrics and slack metrics. The strict indicator is that the entity boundaries and entity types predicted by the model are exactly the same as the standard results. A slack metric is as far as the model predicts entity types that are consistent with the entity types in the standard results, but it is sufficient that the entity boundaries do not overlap with the entity boundaries of the standard results. The set of output results of a given system is denoted as $S=\{s_1,s_2,\cdots,s_m\}$ and the set of standard results (Ground Truth) is denoted as $G=\{g_1,g_2,\cdots,g_n\}$.

(1) The strict index definition $s_i \in S$ is strictly equivalent to $g_t \in G$ and is calculated as in Eqs. (12) and (13):

$$\left(s_i \cdot d_i = g_j \cdot d_j\right) \wedge \left(s_i \cdot b_i = g_j \cdot b_j\right) \wedge \left(s_i \cdot e_i = g_j \cdot e_j\right) \wedge \left(s_i \cdot c_i = g_j \cdot c_i\right) \tag{12}$$

$$P_s = \frac{\left|S \cap_s G\right|}{\left|S\right|}, R_s = \frac{\left|S \cap_s G\right|}{\left|G\right|}, F_{1,s} = \frac{2PR}{P+R} \tag{13}$$

where d_* is the number of the document, b_* and e_* are the start and end positions of the entity mentions in the document, respectively, and c_* is the pre-defined entity category, which defines the strict intersection of the set S and G as $\cap_s$.

(2) The relaxation index defines $s_i \in S$ and $g_t \in G$ as relaxation equivalence, and the relaxation intersection of the set S and G has been defined as $\cap_r$, which is computed as in Eqs. (14) and (15):

$$\left\{\max\left(s_i \cdot b_i, g_j \cdot b_j\right) \le \min\left(s_i \cdot e_i, g_j \cdot e_j\right)\right\} \wedge \left(s_i \cdot c_i = g_j \cdot c_i\right) \tag{14}$$

$$P_r = \frac{|S \cap_r G|}{|S|}, R_r = \frac{|S \cap_r G|}{|G|}, F_{1,r} = \frac{2PR}{P+R} \tag{15}$$

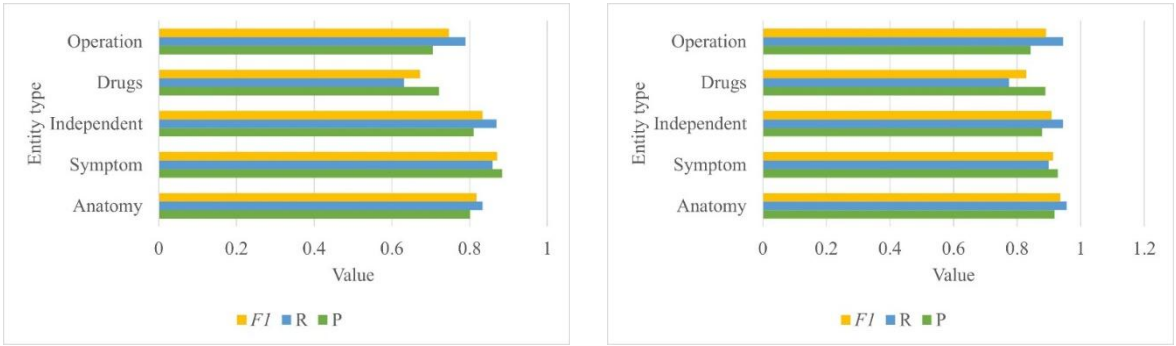
3.1.3 Experimental setup

In the experiments detailed within this chapter, the experimental configuration is composed of two primary components. The first element pertains to the experimental environment arrangement. The entire experiment is conducted within the Ubuntu 16.04 operating system. Python serves as the programming language, and the GPU model employed is the NVIDIA RTX 1080Ti.Regarding the network parameters, the initial learning rate for the entire network is set at 0.003. The dimensionality of the word vectors is 100. The batch size is 32, and the Dropout rate stands at 0.5. For the training process, the Adam optimization algorithm is utilized for 50 epochs.

3.1.4 Model comparison experiments

On the CCKS 2018 dataset, the MUCT model is juxtaposed with BiLSTM and BiLSTM - CRF, i.e. In the training phase, the experiments were set up with essentially the same structure as the original network. All data from the Resume dataset were input into the BiLSTM network for pre-training, and some data from the Resume dataset were randomly selected for testing. When the test *F1* values stabilized, the model parameters with the highest *F1* were selected as initialization parameters, followed by retraining and testing the network model using the CCKS 2018 dataset.

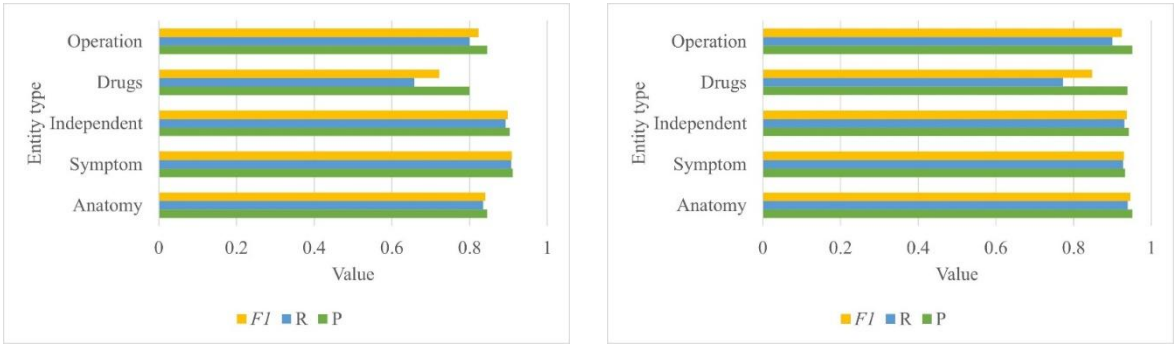
The experimental results of the BiLSTM model on the CCKS 2018 dataset are shown in Figure 3. (a) and (b) respectively represent the precision, recall, and *F1* value under the "strict" and "loose" indicators (the same below). Since the softmax classifier only considers the probability of the labels, it ignores the context, word order, and semantic information in the sentence, resulting in low "strict" *F1* and "loose″ *F1* values for the five categories. The recognition effect of drugs is the worst, with only a "strict" *F1* value of 67.25%.



(a) Strict (b) Relaxed

Figure 3: BiLSTM model experimental results

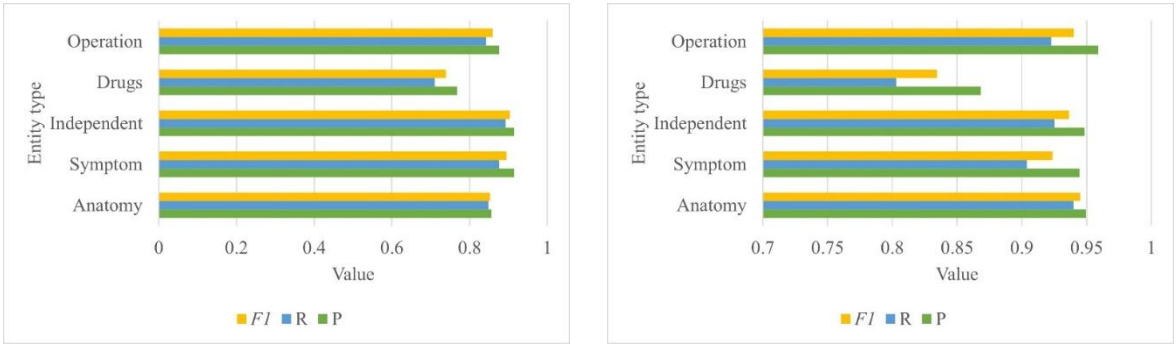
Figure 4 presents the outcomes of the experiments conducted on the CCKS 2018 dataset using the BiLSTM - CRF model. From the perspective of categories, compared with BiLSTM, the BiLSTM-CRF model uses the CRF layer as the output, and its "strict" *F1* and "relaxed" *F1* have significantly improved. However, the results for drug and surgical entity recognition are relatively low. The "strict" *F1* values are 72.2% and 82.23%, while the "relaxed" *F1* values are 84.76% and 92.42%.



(a) Strict (b) Relaxed

Figure 4: BiLSTM-CRF model experimental results

Figure 5 presents the comprehensive experimental outcomes of the MUCT model put forward in this article for the identification of five categories of medical entities. From the perspective of categories, this model performs well in symptom descriptions and independent symptoms. Its "strict" *F1* and "loose" *F1* are 89.46% and 92.35%, 90.33% and 93.65%, respectively. Compared with the BILSTM-CRF model, the "strict" F1 of this model in terms of drug and surgical entities has increased by 1.64% and 3.6%, respectively. In terms of surgery and anatomical sites, the "loose" indicators have higher *F1* values, which are 94.01% and 94.49%, respectively. However, the *F1* values are lower in the "strict" indicators, which are 85.81% and 85.2%, respectively. This indicates that the boundaries of surgery and anatomical sites are prone to recognition errors. In addition, the "strict" *F1* of drugs is the lowest. Due to the majority of drug records in clinical electronic cases being in uppercase letters or abbreviations, the model makes recognition errors. With an increase in the training data of drugs or the addition of external dictionaries, the *F1* of drugs may be improved.



(a) Strict (b) Relaxed

Figure 5: MUCT model experimental results

3.1.5 Ablation experiments

To further confirm the efficacy of each element within the proposed MUCT network model, we carried out ablation tests and assessed the subsequent variants:

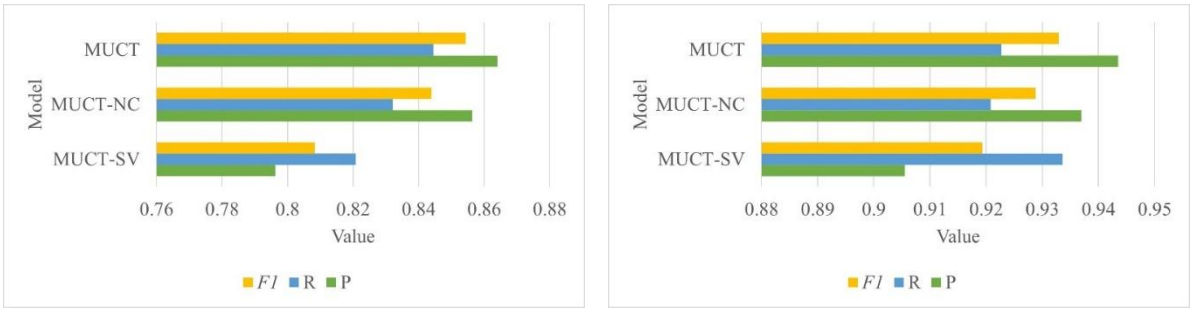
(1) MUCT-SV: using one classifier.

(2) MUCT-NC: removal of the consistency filter.

(3) MUCT: the complete process of the model, containing two employed heterogeneous classifiers with a consistency filter.

The detailed results of the overall P, R and F1 of the ablation experiments under the “strict” and “relaxed” metrics are shown in Fig. 6. By observing the experimental results, it can be

found that the performance of the network with two heterogeneous classifiers and one consistent filter is greatly improved, and its "strict" *F1* is 85.4%, which is 4.57% and 1.08% higher than that of MUCT-SV and MUCT-NC, respectively. This validates the effectiveness of the network pre-training process in migrating knowledge and improves the learning ability of the network model. This is because the cross-domain migrated knowledge plays an important role in guiding the training of the target network. The Resume dataset selected for pre-training is also based on the named entity recognition task, and the data contains eight entity types, such as person's name, country, title, education, and organization. Therefore the training goal of the pre-training phase and the tuning process in the next phase are basically the same. In the tuning process, the MUCT model combines the characteristics of the CCKS 2018 dataset and readjusts the parameter distribution, which makes the extracted features generalization and expressive ability enhanced.



(a) Strict (b) Relaxed

Figure 6: The integrity of the ablation experiment can be compared

3.2 Typical Application Case Analysis

3.2.1 Corpus sources

In this paper, we use the text of online questions from lung cancer patients provided by Dr. A as the source domain corpus, totaling 11,855 items. This dataset comes from real user questions in the lung cancer community of the online medical Q&A platforms Seeking Medicine, Quick Ask Doctor and Good Doctor Online, which have been filtered and labeled with entity names, question types, and entity relationships in detail. To ensure the quality of migration and the authenticity of the study data, the target domain data were also crawled from online medical Q&A websites using Python crawler tool to extract real patient questions, and the detailed data sources are shown in Table 3. The dataset contains a total of 9367 question text counts.

Table 3: The primary corpus of hepatocellular carcinoma patients

Online medical consultation platform	Question text number
Ask the doctor quickly	4563
Good doctor online	2541
Medical information platform	2263
Total	9367

3.2.2 Results of corpus labeling

After formal annotation and manual verification, 1,800 annotated health question texts for liver cancer patients were obtained, with a total of 10,452 entities annotated. The distribution of the nine labels of body site (T1), cellular entity (T2), diagnostic procedure (T3), medication (T4),

metric (T5), individual (T6), question (T7), treatment procedure (T8), and cancer stage (T9) is shown in Figure 7. The problem (T7) label was the most common, accounting for 47.1% of the total, followed by the body site (T1), individual (T6), cancer staging (T9), and treatment procedure (T8) labels. The remaining labels were more evenly distributed and labeled with a relatively small amount of entities, such as diagnostic procedures (T3), medications (T4), metrics (T5), and cellular entities (T2) labels.

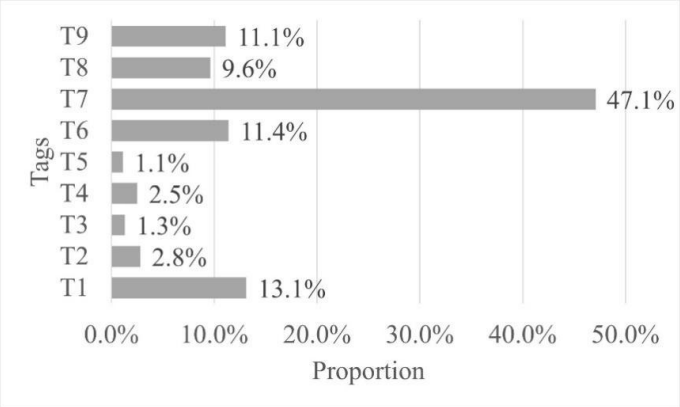


Figure 7: The health question of patients with liver cancer is distributed

Since the corpus of questions asked by lung cancer patients has been manually labeled with entities and entity relations, the question type annotation in it was deleted and the entity annotation part was retained for the purpose of applying it to this study. The entity annotation part deleted some tags such as time, frequency, food, geographic location, and chemical substance according to the annotation framework of this paper. Age and demographic labels were merged to unify the labeling with individual labels. Finally, 11,843 annotated health question texts for liver cancer patients were obtained, and a total of 46,857 entities were labeled. The specific label distribution is shown in Figure 8. From the perspective of the difference in label distribution, it can be seen that most of the health questions of liver and lung cancer patients mentioned relatively exact descriptions of their conditions, i.e., the entities reflected in the labels of the question (55.3%). The reason for the more even distribution of body parts (14.4%) and individual (5.5%) labels and the larger number of labeled entities may be related to the requirements of the question format of the relevant health websites. For example, the question-and-answer section of Seeking Medical Help.com provides patients with a quick channel to ask questions by type of disease, and requires patients to fill in personal information such as main symptoms, changes in condition, and age and gender in a step-by-step manner, which explains to a certain extent the distribution of the number of entities labeled with questions, body parts, and individuals. As can be seen from the relatively small percentage of medication (3.2%) and metric (0.7%) labels, only a few health questioners in the available research corpus made descriptions related to medication name, medication use, or medication dosage during the questioning process. Furthermore, as this study mainly focuses on the health issues of cancer patients, the newly added cancer staging labels in the annotation system also demonstrate their significance. The marking recognition includes "early stage", "mid-stage", "late stage", as well as more precise TNM staging, such as "primary liver cancer T4N1M1IIIB stage". The identification of cancer staging helps information users to have a more accurate understanding of the stage of the cancer patient's condition, thereby improving the efficiency of using the patient's question information.

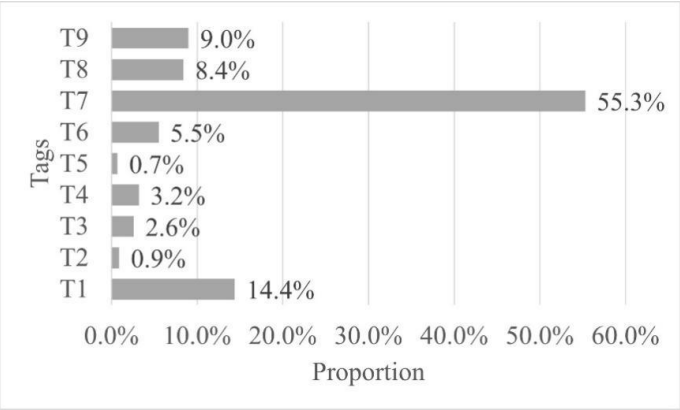


Figure 8: The health questions of lung cancer patients are distributed

The case study results show that the migration method only needs to utilize a small amount of annotated corpus in the target domain to effectively identify entities such as personal information, disease symptoms, diagnosis and treatment, and drug usage in the text of patients' questions in this domain, which realizes the full utilization of the existing data resources, and at the same time provides references to the related disease research and natural language processing research.

4 Conclusion

In this paper, a deep migration learning model based on multi-source domain integration (MUCT) is proposed to improve the traditional cross-domain migration learning method, and the following conclusions are drawn from the model performance test and typical case analysis:

(1) The study performs named entity recognition on the CCKS 2018 review dataset, and the strict *F1* value of the MUCT model on the CCKS 2018 review dataset is improved by 4.57% and 1.08% compared with MUCT-SV and MUCT-NC, respectively, which verifies the migration learning effect of the MUCT model.

(2) A corpus annotation framework for online questioning texts of liver cancer patients was established by selecting nine entity tags that are closely related to clinical diagnosis. By migrating the questioning text of liver cancer patients, the corpus annotation results show that the MUCT model only needs to utilize a small amount of annotated corpus in the target domain to effectively analyze the biological data, which verifies that the MUCT model has an efficient transfer learning effect.

However, the experiments in this paper only migrated the text of liver cancer patients' questions, and did not involve other factors such as disease types or data sources. In order to further solve the domain adaptation problem, more data from different data sources and disease types can be introduced in the future to validate the model performance, and cross-domain migration experiments can be conducted for disease types with large domain differences to provide references for related disease research and natural language processing research.

Funding

This research was supported by the Clinical Medicine Research Special Project (2024LYA05).

References

[1] Samdarshi, S. (2024). Genome sequencing and bioinformatics. Frontiers in Molecular Genetics and Genomics, 162-187.

[2] Fu, Y., Ling, Z., Arabnia, H., & Deng, Y. (2020). Current trend and development in bioinformatics research. BMC bioinformatics, 21(Suppl 9), 538.

[3] Broekema, R. V., Bakker, O. B., & Jonkers, I. H. (2020). A practical view of fine-mapping and gene prioritization in the post-genome-wide association era. Open biology, 10(1).

[4] Lee, J. Y. (2023). The principles and applications of high-throughput sequencing technologies. Development & Reproduction, 27(1), 9.

[5] Guo, M., & Zou, Q. (2019). Perspectives of bioinformatics in big data era. Current genomics, 20(2), 79.

[6] Wolf, F. A., Angerer, P., & Theis, F. J. (2018). SCANPY: large-scale single-cell gene expression data analysis. Genome biology, 19(1), 15.

[7] Dall'Alba, G., Casa, P. L., Abreu, F. P. D., Notari, D. L., & de Avila e Silva, S. (2022). A survey of biological data in a big data perspective. Big Data, 10(4), 279-297.

[8] Nagaraj, K., Sharvani, G. S., & Sridhar, A. (2018). Emerging trend of big data analytics in bioinformatics: a literature review. International Journal of Bioinformatics Research and Applications, 14(1-2), 144-205.

[9] López-Fernández, A., Gomez-Vela, F. A., Rodriguez-Baena, D. S., Delgado-Chaves, F. M., & Gonzalez-Dominguez, J. (2025). Biclustering in bioinformatics using big data and High Performance Computing applications: challenges and perspectives, a review: A. Lopez-Fernandez et al. The Journal of Supercomputing, 81(10), 1123.

[10] Li, Y., Wu, F. X., & Ngom, A. (2018). A review on machine learning principles for multi-view biological data integration. Briefings in bioinformatics, 19(2), 325-340.

[11] Gharajeh, M. S. (2018). Biological big data analytics. In Advances in computers (Vol. 109, pp. 321-355). Elsevier.

[12] Mirza, B., Wang, W., Wang, J., Choi, H., Chung, N. C., & Ping, P. (2019). Machine learning and integrative analysis of biomedical big data. Genes, 10(2), 87.

[13] Gilpin, W., Huang, Y., & Forger, D. B. (2020). Learning dynamics from large biological data sets: machine learning meets systems biology. Current Opinion in Systems Biology, 22, 1-7.

[14] Pan, S. J. (2020). Transfer learning. Learning, 21, 1-2.

[15] Rogers, A. W., Vega-Ramon, F., Yan, J., del Río-Chanona, E. A., Jing, K., & Zhang, D. (2022). A transfer learning approach for predictive modeling of bioprocesses using small data. Biotechnology and Bioengineering, 119(2), 411-422.

[16] Theodoris, C. V., Xiao, L., Chopra, A., Chaffin, M. D., Al Sayed, Z. R., Hill, M. C., ... & Ellinor, P. T. (2023). Transfer learning enables predictions in network biology. Nature, 618(7965), 616-624.

[17] Ashayeri, H., Sobhi, N., Pławiak, P., Pedrammehr, S., Alizadehsani, R., & Jafarizadeh, A. (2024). Transfer learning in cancer genetics, mutation detection, gene expression analysis, and syndrome recognition. Cancers, 16(11), 2138.

[18] Zhang, S., Zhou, Y., Dong, K., Liu, J., Geng, P., & Lu, Q. (2025). Predictive modeling and inference using deep transfer learning in genetic data analysis. Statistics Innovation, 2(1).

[19] Muneeb, M., Feng, S., & Henschel, A. (2022). Transfer learning for genotype–phenotype prediction using deep learning models. BMC bioinformatics, 23(1), 511.

[20] Giorgi, J. M., & Bader, G. D. (2018). Transfer learning for biomedical named entity recognition with neural networks. Bioinformatics, 34(23), 4087-4094.

[21] Kensert, A., Harrison, P. J., & Spjuth, O. (2019). Transfer learning with deep convolutional neural networks for classifying cellular morphological changes. SLAS Discovery: Advancing Life Sciences R&D, 24(4), 466-475.

[22] Pickering, A., Martinez Balvanera, S., Jones, K. E., & Hedwig, D. (2025). A scalable transfer learning workflow for extracting biological and behavioural insights from forest elephant vocalizations. Remote Sensing in Ecology and Conservation.

[23] Zheng, M., Yang, B., & Xie, Y. (2020). EEG classification across sessions and across subjects through transfer learning in motor imagery-based brain-machine interface system. Medical & biological engineering & computing, 58(7), 1515-1528.

[24] Bird, J. J., Kobylarz, J., Faria, D. R., Ekárt, A., & Ribeiro, E. P. (2020). Cross-domain MLP and CNN transfer learning for biological signal processing: EEG and EMG. IEEe Access, 8, 54789-54801.

[25] Gu, Y., Ge, Z., Bonnington, C. P., & Zhou, J. (2019). Progressive transfer learning and adversarial domain adaptation for cross-domain skin disease classification. IEEE journal of biomedical and health informatics, 24(5), 1379-1393.

[26] Guo, C., Wei, B., & Yu, K. (2021). Deep Transfer Learning for Biology Cross-Domain Image Classification. Journal of Control Science and Engineering, 2021(1), 2518837.

[27] Yan, K., Guo, X., Ji, Z., & Zhou, X. (2021). Deep transfer learning for cross-species plant disease diagnosis adapting mixed subdomains. IEEE/ACM transactions on computational biology and bioinformatics, 20(4), 2555-2564.

[28] Maswanganyi, R. C., Tu, C., Owolawi, P. A., & Du, S. (2023). Multi-class transfer learning and domain selection for cross-subject EEG classification. Applied Sciences, 13(8), 5205.

[29] Li, Q., Wang, Y., Zhang, Y., Zuo, Z., Chen, J., & Wang, W. (2024). Fuzzy-vit: A deep neuro-fuzzy system for cross-domain transfer learning from large-scale general data to

medical image. IEEE Transactions on Fuzzy Systems, 33(1), 231-241.