Generalizability of Concept Knowledge in Machine Learning Using TCAV Scores: A Case Study Using Different Skin-Lesion Datasets

Moritz C. Schwinghammer*†, Laines Schmalwasser†, Sireesha Chamarthi†, Yuri A.W. Shardt*

**Technical University Ilmenau: Department of Automation Engineering, Ilmenau, 98684, Germany*
†German Aerospace Center: Institute for Data Science, Jena, 07745, Germany

Abstract: In safety-critical fields, such as skin-lesion classification, interpretability of the decisions of a machine learning model is required. This can be provided through concept-based interpretability methods like testing with concept activation vectors (TCAV). TCAV quantifies how specific human-understandable concepts influence a model's decisions. A further issue affecting the performance of ML models is generalizability, *i.e.*, how well a model generalizes to unseen data from a different domain. It is currently unknown how the interpretability provided by TCAV is affected by domain shifts. Here we show that TCAV-based interpretability is predominantly unaffected by domain shifts. To that end, we introduce concept detection scores (CDS) as aggregated TCAV scores which are directionally unified and thus a suitable evaluation metric. The results show only small differences between CDS within domain and across domain for 48 models trained on three distinct source domains. This increases the viability of TCAV as an interpretability tool since it can be used without additional effort to manage generalizability.

Keywords: **Interpretability, TCAV, Generalizability, Skin-Lesion Classification, Domain Shifts**

1. INTRODUCTION

The adoption of artificial intelligence, particularly in safety-critical areas, has been slow, partly due to the presence of uncertainties in AI systems and the prevalent distrust of the opaque decision-making processes in neural-network (NN) models [1]. These concerns are addressed by explainable artificial intelligence (xAI), which seeks to provide interpretability by opening the machine learning (ML) black box and explaining the models' decisions.

Various approaches to xAI exist, varying in temporal type (*post-hoc* or *ante-hoc*), in the nature of the explanations provided (from prototypes to decision-governing rules), and in the scope of explanations (from explaining individual samples to global interpretability for a NN) [2]. One area of interpretability is concept-based NN explanations. These explanations focus on explaining the factors that lead to a NN model's decisions in human-understandable terms. A prominent, global, *post-hoc* interpretability method that identifies representation vectors of human-understandable concepts in the activation (latent) space of a NN and quantifies their influence on the model's predictions, is testing with concept activation vectors (TCAV) [3].

Skin-lesion classification (SKL) is a safety-critical, high-risk field where dermatologists often seek a "qualified second opinion" from another expert [4]. If the second opinion is provided by computer-aided diagnostics (CAID), it must be able to explain its reasoning, as a dermatologist would, rather than simply giving a final classification [4].

Concept-based interpretability tools, such as TCAV, offer one xAI approach for decision reasoning. This is because dermatologists themselves rely on specific criteria or concepts to classify skin lesions. Those criteria are encoded in several checklists, which in turn provide guidance for SKL depending on the presence, absence, or type of observable criteria on a lesion [5]. Thus, as [6, 7] have shown, concept-based interpretability of a CAID application's decisions can use the same terms and criteria, essentially the same "language," as the experts it assists.

However, it is currently unclear how this interpretability method handles domain shifts, especially since ML models in general have issues with generalizability. When a ML model is given unseen data from a distribution (domain) different from the one on which it was originally trained, then its predictive performance often degrades significantly [8, 9]. In this case, "different domain" refers not merely to an unseen test-data split, but to a fundamentally different distribution of a dataset's inherent features [8]. A relevant example of a technical domain shift between different skin-lesion (SK) datasets would be image-acquisition system settings like contrast and brightness, which differ between the datasets, since the images were captured in different hospitals [9]. Addressing the generalizability issues is a major concern in ML-based SKL [9].

Therefore, before CAID tools based on concept-based explanations can be applied in real-world scenarios, it is necessary to assess how these explanations handle domain shifts, *i.e.*, when the model must extrapolate to new domains. In this paper, TCAV and the newly introduced concept detection scores (CDS), which aggregate TCAV scores to address dimensionality and directional inconsistency hindering the evaluation, will be used to provide an understanding of the degree to which different ML models learn and apply concepts across three datasets. By comparing concept knowledge from a known domain with concept knowledge on an unknown domain, conclusions about

generalizability can be drawn. The results show that, for the NNs used, concept knowledge is largely unaffected by technical domain shifts, suggesting robustness in concept-based explanations for xAI in CAID applications.

2. THEORY AND BACKGROUND

2.1. Dermatological concepts

Although studies show that early detection of skin lesions is challenging, dermatologists use heuristic approaches, such as the seven-point checklist, to diagnose and encode their expertise [5]. The diagnoses used in this paper are either *nevus* or *melanoma*. The concepts used in this paper are derived from the criteria of the seven-point checklist [5] and described in accordance with [6, 10] as:

1) Pigment Networks (PN): A pigment network is a grid-like pattern of interconnected lines surrounding lighter areas. A typical pigment network (PN_T), which indicates a benign lesion, is symmetrical and consistent. An atypical pigment network (PN_AT) is asymmetrical, with variable color, thickness, and spacing, and suggests the presence of *melanoma*.
2) Blue-Whitish Veil (BWV): This concept refers to an irregularly shaped, slightly blue lesion (spot) covered by a whitish haze resembling ground glass.
3) Streaks (ST): Streaks can be either regular (ST_R), indicating a benign lesion, or irregular (ST_IR), suggesting melanoma. In general, they appear as straight extensions, bulbous projections, or a widened network along the lesion edge.
4) Dots and Globules (DG): Regular dots and globules (DG_R) are centered within the lesion middle or aligned on the network lines and are uniform, indicating a benign lesion. Irregular dots and globules (DG_IR) exhibit higher variability, suggesting melanoma.
5) Regression Structures (RS): The presence of fine grey-bluish dots, light areas without blood vessels, or shiny-white structures indicates regression structures, suggesting melanoma.

2.2. Concept detection and testing with concept activation vectors

Testing with concept activation vectors (TCAV), introduced by Been Kim *et al.* [10], is a method for interpreting neural networks by analyzing concept relevance. TCAV involves splitting a NN at a specified layer l with e neurons. The model m described by $g_m(x)\colon \mathbb{R}^a \to \mathbb{R}^k$, maps inputs $x \in \mathbb{R}^a$ to a logit space $\mathbb{R}^k$ where x represents the pixel-based input images. The activations at layer l, denoted by $o_{l,m}(x)\colon \mathbb{R}^a \to \mathbb{R}^e$, are used to train a binary classifier for each labeled individual concept c with $c \in C$ and C being the set of all observed concepts. The normal to the hyperplane of the linear classifier is the concept activation vector (CAV), $v^l_{c,m}$, which points toward the concept encoded in the e-dimensional activation space $\mathbb{R}^e$.

To compute TCAV scores, input data must have classification class labels k, corresponding to the class predicted by the NN's final (logit) layer. The "classifying" part of the NN complements the "feature extracting" NN part $o_{l,m}(x)$ and is described with $h_{l,m} : \mathbb{R}^e \to \mathbb{R}^k$. The sensitivity $S^{Sens}_{c,k,l,m}(x)$ measures how sensitive a model m is to a concept c for classification class k at layer l. It is described as the directional derivative:

$$S^{Sens}_{c,k,l,m}(x) = \lim_{\varepsilon \to 0} \frac{h_{l,k,m}\left(o_{l,m}(x) + \varepsilon v^l_{c,m}\right) - h_{l,k,m}\left(o_{l,m}(x)\right)}{\varepsilon}, \quad (2.1)$$

$$S^{Sens}_{c,k,l,m}(x) = \nabla h_{l,k,m}\left(o_{l,m}(x)\right) \cdot v^l_{c,m},$$

with $S^{Sens}_{c,k,l,m}(x)$ providing quantifiable information about the sensitivity of the classification of a single input image x. For global *post-hoc* interpretability, TCAV scores aggregate S^{Sens} globally, thus quantifying the relevance of a concept c across all samples belonging to a classification class k. The TCAV scores are defined as:

$$S^{TCAV}_{c,k,m,w} = \frac{\left|\{x \in X_k : S^{Sens}_{c,k,l,m}(x) > 0\}\right|}{|X_k|}, \quad (2.2)$$

which is the fraction of all predictions of X_k where the concept was classification-relevant, over all images of X_k, the set containing all samples of a specific class k [3]. To account for variations in data preprocessing and classifier initialization, S^{TCAV} is computed in multiple runs ($w > 1$) using different samplings of the classifier training set [3].

2.3. Concept detection scores

While S^{TCAV} are averaged S^{Sens}, there are still $|C||K|$w individual TCAV scores per model m with $c \in C$ and $k \in K$, which due to the different dependencies presents a challenge for analysis. Here, the set C is the set containing all concepts c and K is the set containing all classes k. The multitude of TCAV scores, as well as the issue of directionality, motivated the introduction of concept detection scores (CDS).

The issue of directionality refers to the fact that depending on concept c and classification class k, "perfect" concept detection results in either a high S^{TCAV} of 1 or a low S^{TCAV} with a value of 0. This is due to the fact that the concepts c are all indicative of precisely one of the classification classes k. Thus, if S^{TCAV} is calculated for a concept c which is indicative of another classification class k than the one examined during calculation of the S^{TCAV}, a perfect concept detection would result in a score of 0. This inversion, which depends on the combination of concept c and classification class k, poses a significant hindrance for later analysis of concept knowledge.

To address this, CDS serve two purposes. First, the reduction of the dimensionality of the S^{TCAV} results through aggregation. Second, a partial inversion of the calculated S^{TCAV} to norm the direction of the detected concept knowledge, with 1 being always indicative of a "perfect" concept detection.

The reduction in dimensionality is feasible because all observed concepts c are indicative of one element of the binary class k. This allows aggregation of all concepts for each element of class k, which is controlled by the variable concept type group B defined as:

$$B = \begin{cases} b_1 & c \in C_{negative} \\ b_2 & c \in C_{positive} \\ b_3 & c \in C \end{cases}, \quad (2.3)$$

with $C_{negative}$ and $C_{positive}$ being the sets of all concepts c indicative of either classification class k. While B accounts for the indicative k of the individual c, CDS are also dependent on the basic k, since the underlying S^{TCAV} depend on it. For the purpose of CDS, k is grouped into the target class group D, defined as:

$$D = \begin{cases} d_1 & S^{TCAV}_{c,k_{negative},m,w} \\ d_2 & S^{TCAV}_{c,k_{positive},m,w} \\ d_3 & d_1 \cup d_2 \end{cases}. \quad (2.4)$$

The individual CDS are described:

$$S^{CD}_{b_1,d_1} = \frac{\sum_{c_{neg}}^{|C_{negative}|} \sum_{i=0}^{w} S^{TCAV}_{w,k_{negative},c_{neg}}}{|w| \cdot |C_{negative}|} \quad (2.5)$$

$$S^{CD}_{b_1,d_2} = \frac{\sum_{c_{neg}}^{|C_{negative}|} \sum_{i=0}^{w} 1 - S^{TCAV}_{w,k_{positive},c_{neg}}}{|w| \cdot |C_{negative}|} \quad (2.6)$$

$$S^{CD}_{b_2,d_1} = \frac{\sum_{c_{pos}}^{|C_{positive}|} \sum_{i=0}^{w} 1 - S^{TCAV}_{w,k_{negative},c_{pos}}}{|w| \cdot |C_{positive}|} \quad (2.7)$$

$$S^{CD}_{b_2,d_2} = \frac{\sum_{c_{pos}}^{|C_{positive}|} \sum_{i=0}^{w} S^{TCAV}_{w,k_{positive},c_{pos}}}{|w| \cdot |C_{positive}|} \quad (2.8)$$

where c_{neg} and c_{pos} are individual concepts indicative of either $k_{negative}$ or $k_{positive}$.

3. METHODOLOGY

3.1. Datasets

Three types of SK image classification datasets were used in the paper. The first type, the *concept dataset,* consists in this case of a single dataset containing the concept information used to train the CAVs. This dataset includes SK images with labels for both classes k, as well as the presence or absence of concepts. The second dataset type, referred to as the *random dataset*, also consists of a single dataset from which images were randomly selected and paired with random concept labels. The final dataset type comprises three *domain-shifted datasets*. Domain shifts mean that they contain different distributions (domains) which have observable shifts in their features and/or characteristics [9].

The *concept dataset* used was the Seven-point Criteria Evaluation Database defined as f^{cnc}_{d7pt} [10]. The concepts, the classification class k they indicate, and the number of images containing each concept are shown in Table 1.

Table 1: Overview of all concepts applied

Concept of seven-point checklist	Pheno-type	Abbre-viation	Indicates k	# of imgs
Pigment network	typical	PN_T	~~*mel*~~	335
	atypical	PN_AT	*mel*	216
Blue whitish veil		BWV	*mel*	183
Streaks	regular	ST_R	~~*mel*~~	96
	irregular	ST_IR	*mel*	237
Dots and globules	regular	DG_R	~~*mel*~~	301
	irregular	DG_IR	*mel*	392
Reg. structures		RS	*mel*	183

"Indicates k" with a value of ~~*mel*~~ refers to the fact that the presence of said concept indicates the absence of *melanoma*

The *random dataset* was the train split of the ISIC2018 dataset [11]. Finally, the *domain-shifted datasets* were datasets BCN20000, HAM10000 and MSK with 2721, 4234 and 1282 SK images with class k_{nevus} and 1918, 465 and 565 SK images for class $k_{melanoma}$ respectively. These datasets were separated by [9] into their underlying domains, which could be both technical, as well as biological. This paper focuses on the technical domains which are in this case the technical differences between the datasets such as properties of the image-acquisition system used in the datasets' creation [9].

3.2. Procedure

In accordance with [6], concept training was repeated twenty times on each NN model for each concept to guard against statistical influence on classifier capability, thus setting $w = 20$. Consequently, this required twenty individual stratified splits of f^{cnc}_{d7pt} for each combination of $S^{TCAV}_{c,k,m,w}$ variables. These splits were further split into cluster-based undersampled train splits, and as-is evaluation splits for each individual concept c. The per-concept-balanced training splits were then used to train twenty individual logistic regression classifiers per concept for each NN model. The training splits were used as input and passed through the NN models until reaching the split layer l, where activations were extracted. Across all NN models, regardless of the model architecture, layer l was consistently set as the first regularization or flattening layer immediately following the final imported convolutional layer of the primary model architecture (VGG16, VGG19, InceptionNetV3, or ResNet50).

Training w separate classifiers helped reduce randomness and ensured the robustness of the TCAV results. Additionally, as part of the process, a random baseline was calculated to serve as a comparison point. Using randomly selected images from the *random dataset* as input verified whether the calculated results were significantly different from those produced by classifiers trained on data without useful information. To achieve this, the *random dataset* ISIC2018 was used as the source for the randomly selected images, and defined as $f^{rnd}_{ISIC2018}$. The images of $f^{rnd}_{ISIC2018}$ were then assigned a partially random label, either k_{nevus} or $k_{melanoma}$. In accordance with the literature, $k_{melanoma}$ is the positive class or $k_{positive}$, since in SKL it is imperative to detect all *melanoma* cases. The class distribution of the real per-concept-balanced train splits which were not equal because of the cluster-based undersampling was mirrored with the randomly assigned labels. The total number of classifiers trained on random images from $f^{rnd}_{ISIC2018}$ were 160 as determined by $w|C| = 160$, since eight individual concepts c comprise set C. All 160 random baseline classifiers, along with the corresponding TCAV scores, formed the population for a single random concept.

The classifier training process began with extraction of activations from activation space $\mathbb{R}^e$ of the NN model for each per-concept-balanced training split. The activations, together with their corresponding labels, were then shuffled and passed to a linear classifier. The type of classifier used for the results presented in the paper was a linear classifier with L_2 regularization. For evaluation of the classifiers, the per-concept-balanced test splits were fed as input into the primary NN model, where the activations were extracted at layer l,

before being fed into the trained classifiers. The performance of each classifier was then tracked using the metrics: accuracy, precision, recall, and F1 score.

The process of model training, classifier training and then concept knowledge detection on the concept target datasets f^{tcn} is shown in Figure 1. Concept target datasets f^{tcn} are the *domain-shifted datasets* upon which the concept knowledge of the models is evaluated. For each model architecture one m was trained on each of the three *domain-shifted datasets*, defined as f^{src}. Each of the three f^{tcn} was then evaluated against every m, resulting in one within-domain evaluation and two cross-domain evaluations. However, for concept detection on f^{tcn}, the CAV first had to be determined, *i.e.*, the coefficient vector for each classifier was extracted as the concept activation vector $v^l_{c,m}$. Multiplication of the dot product of $v^l_{c,m}$ and the gradient of the classification part of the model $\nabla h_{l,k}(o_l(x))$ resulted in the sensitivity $S^{Sensitivity}_{c,k,l}(x)$ per image x. Aggregation of all sensitivities using TCAV scores resulted in w-times S^{TCAV} for each individual concept c and for both classification classes k_{nevus} and $k_{melanoma}$.

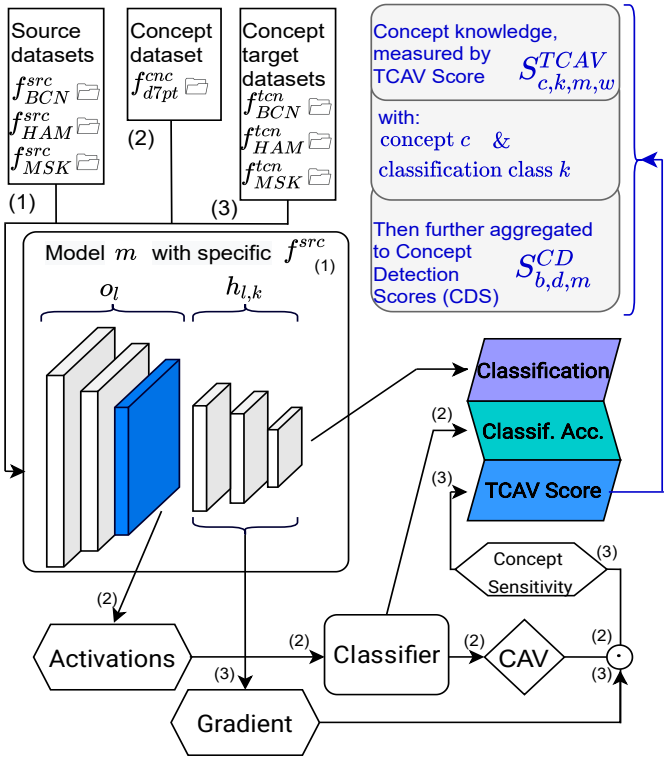

Figure 1: Process overview for a single model m, showing (1) model training on source f^{src}, (2) classifier training and evaluation, and (3) concept knowledge evaluation

4. RESULTS AND ANALYSIS

Before exploring concept detection on unlabeled *domain-shifted datasets*, it is necessary to first validate the concept detection on labeled test splits of f^{cnc}_{d7pt} against the baseline. The validation (or random) baseline is calculated using randomly selected images from the dataset $f^{rnd}_{ISIC2018}$ and is expected to show no bias towards the presence or absence of concepts [9]. Since the results in this paper range between 0 and 1, the baseline is expected to be around 0.5. Our results of the calculation of "random classifiers" on a single model m corroborate this observation, with the mean F1 score being 0.510 and the mean accuracy being 0.505. Thus, the classifier baseline can be set to 0.5, in accordance with expectations and observed results. A larger difference would indicate a hidden bias in the model.

Since none of the concept target datasets f^{tcn}_{BCN}, f^{tcn}_{HAM}, and f^{tcn}_{MSK} provide concept labels, the test splits of the *concept dataset* f^{cnc}_{d7pt} are used for validation. A general overview of the fidelity of concept detection through the classifiers on f^{cnc}_{d7pt} is presented in Figure 2. The figure confirms that all concepts, regardless of which class k they indicate, are detected by the classifiers with an average total F1 score of 0.705. Moreover, almost all classifiers correctly recognize their respective concept c. Out of the 960 classifiers trained per c across all NN models, fifteen classifiers for concept BWV, six classifiers for concept PN_AT, and four classifiers for concept ST_R have F1 scores below the baseline of 0.5. Thus, since most classifiers perform better than the baseline, the general detectability of concept knowledge is validated.

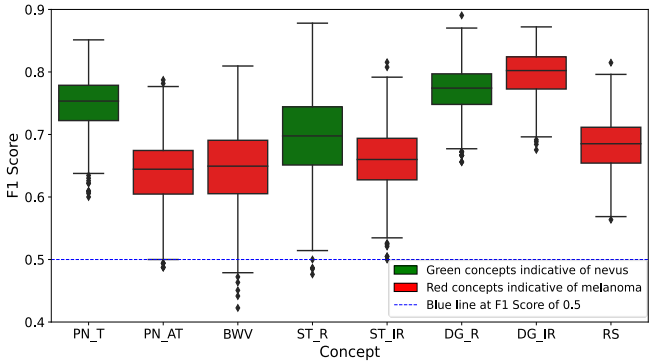

Figure 2: F1 scores of all twenty classifiers per concept c, calculated on the unseen test split of concept dataset f^{cnc}_{d7pt}

With the validation of the concept detection on labeled f^{cnc}_{d7pt} concluded, the concept detection on the unlabeled f^{tcn} can be examined. Here, too, a random baseline for TCAV scores is expected at around 0.5 [6]. This is confirmed by the calculation of 160 TCAV scores for each of the classes k, which resulted in a mean of 0.51 for k_{nevus} and 0.48 for $k_{melanoma}$. Thus, for TCAV as well, the baseline can be set to 0.5. Compared to this baseline, 78.1% of TCAV results are significantly different for a p-value < 0.05.

Individual TCAV scores for a single m, selected for its concept detection capability, with f^{src}_{BCN} and f^{tcn}_{BCN} as source and concept target dataset are shown in Figure 3. Since $f^{src}_{BCN} = f^{tcn}_{BCN}$, the results are within domain. In the left subplot of the figure, it can be observed that $S^{TCAV}_{c_{negative},k_{nevus},m,w}$, are mostly close to 1, while $S^{TCAV}_{c_{positive},k_{nevus},m,w}$ are near zero. This matches the expectations, since it is expected that concepts $c_{negative}$ which are indicative of k_{nevus} are detected and influential on x belonging to $X_{k_{nevus}}$. The signs of most of the TCAV scores per concept, switch for TCAV scores calculated on images belonging to $k_{melanoma}$, as is also expected.

While TCAV scores visualized with mean and violin plots, as in Figure 3, provide a viable gauge of the detected concepts for a single m for the case where $f^{src}_{BCN} = f^{tcn}_{BCN}$, the

comprehensibility decreases greatly when 16 model architectures are trained on all three f^{src}, evaluated on all eight c, and evaluated on all three f^{tcn}. This is where CDS provide an advantage.

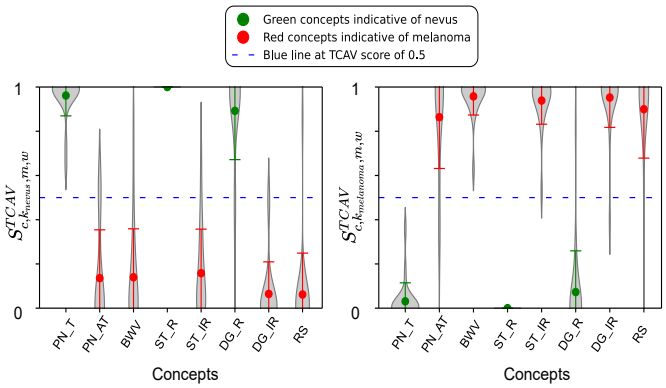

Figure 3: Violin plots of the distribution of TCAV scores S^{TCAV} per concept c of a single m with $f^{src}_{BCN} = f^{tcn}_{BCN}$

This advantage is shown in Figure 4, which shows the concept detection of all models for all S^{CD} and within- and across domain. From top to bottom, the first subplot shows $S^{CD}_{b_1,d_1}$ and thus concepts $c_{negative}$ detected on images of k_{nevus}. The following subplots show in order $S^{CD}_{b_2,d_1}$, $S^{CD}_{b_1,d_2}$, and $S^{CD}_{b_2,d_2}$. Within each subplot, the boxplot color signifies f^{tcn}, while the background color distinguishes f^{src} with f^{tcn}_{d7pt} in neon green as added fourth *concept target dataset*.

The primary observation in Figure 4 is the apparent irrelevance of f^{tcn}**.** Models trained on the same f^{src} seem to report highly similar S^{CD} for all f^{tcn}. This pattern is even more pronounced when only the median is considered, and it consistently applies across all b and d. The highest difference between the medians of the average concept knowledge measured across all f^{tcn} is 0.054 for f^{src}_{BCN} at $S^{CD}_{b_2,d_1}$. The average difference for all f^{src} and S^{CD} is 0.022.

The distribution from Figure 4 and the miniscule differences between CDS of different f^{tcn} both suggest that the specific concept target dataset f^{tcn} is largely irrelevant for the detected concept knowledge and its influence. However, Figure 4 only considers the distribution and median of the CDS. Thus, to observe on the level of individual NN models, Table 2 shows the maximum difference (Δ^{max}) between all instances of f^{tcn} for each individual CDS, m, and f^{src}. The rows with statistic *Max.* report the single highest of Δ^{max} $S^{CD}_{b,d,m}$ per b, d and m, while rows *Median* and *Mean* report the distribution of the Δ^{max} $S^{CD}_{b,d,m}$ across all m The table shows that across all f^{src}, the mean and median Δ^{max} between the instances of f^{tcn} is below 0.056 for the mean and 0.021 for the median, thus further underscoring the observation that f^{tcn} has a negligible impact on the concept knowledge. However, some outliers are also present in Table 2. The total maximum of all *Max.* Δ^{max} across all f^{tcn} shows the highest outliers (value > 0.1) at $S^{CD}_{b_1,d_2,m,f^{src}_{HAM}}$, $S^{CD}_{b_2,d_1,m,f^{src}_{HAM}}$, ~~and~~ $S^{CD}_{b_2,d_2,m,f^{src}_{HAM}}$, and $S^{CD}_{b_1,d_2,m,f^{src}_{MSK}}$. Albeit eight of twelve possible *Max.* Δ^{max} $S^{CD}_{b,d,m}$ are below 0.1, showing that even in most of the worst of the worst-cases, the generalization-induced differences are small. Thus, concept knowledge appears to be largely independent of the dataset on which it is measured, in contrast to predictive capability, which generalizes poorly and is highly domain-dependent.

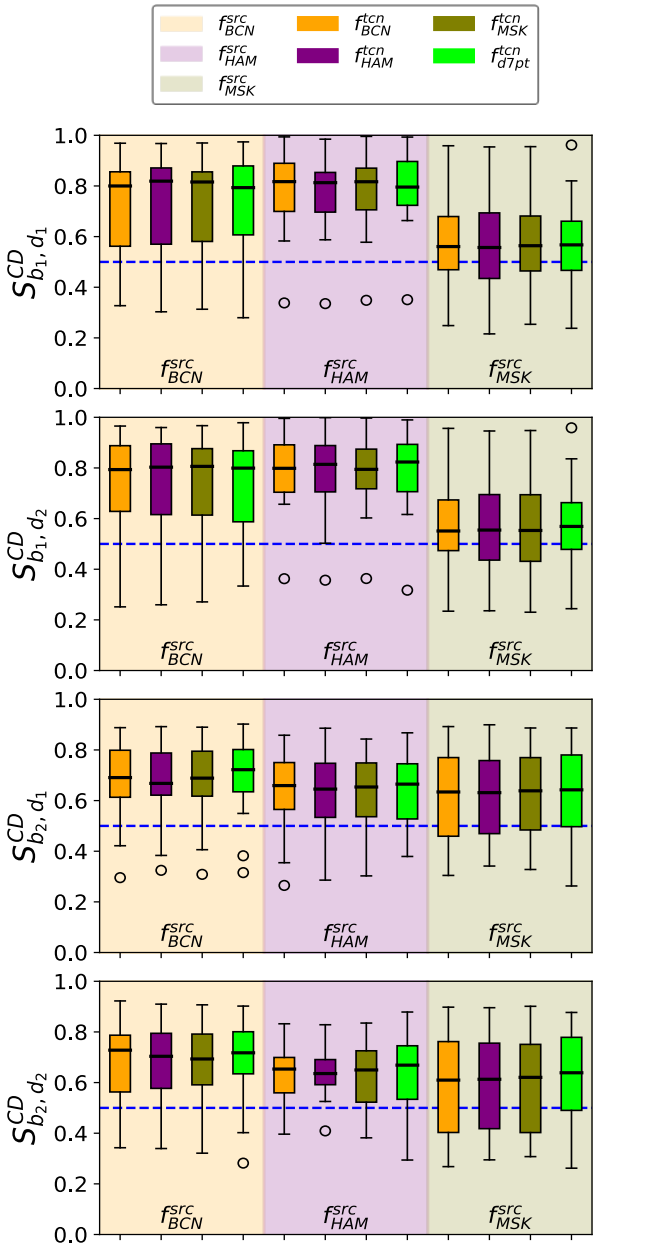

Figure 4: Boxplots of the distribution of concept detection scores $S^{CD}_{b,d}$ for all combinations of b and d separated by source datasets f^{src} and concept target datasets f^{tcn}

The three highest outliers are all trained on f^{src}_{HAM}. However, while f^{src}_{BCN} is overall the best source dataset with respect to (w.r.t.) concept knowledge, as shown in Figure 4, f^{src}_{HAM} is the dataset with the smallest spread between its CDS. Lastly, NN models trained on f^{src}_{MSK} command the least concept knowledge, they consistently have the lowest median CDS across all b and d, particularly for b_1, where the concepts indicate k_{nevus}. This suggests that the concepts in general are harder to learn on f^{src}_{MSK}, especially for c_{nevus}, although learning is still possible.

Table 2: Maximum difference between lowest and highest $S^{CD}_{b,d,m,f^{tcn}}$ between f^{tcn} of individual m for all b, d, and f^{src}

f^{src}	Statistic	Δ^{max} of $S^{CD}_{b_1,d_1,m}$	Δ^{max} of $S^{CD}_{b_1,d_2,m}$	Δ^{max} of $S^{CD}_{b_2,d_1,m}$	Δ^{max} of $S^{CD}_{b_2,d_2,m}$
BCN	Max.	*0.063*	*0.047*	*0.052*	0.096
BCN	Median	0.018	*0.020*	*0.019*	**0.026**
BCN	Mean	0.022	*0.021*	*0.021*	0.031
HAM	Max.	*0.063*	**0.238**	**0.141**	**0.226**
HAM	Median	*0.013*	**0.029**	0.023	0.021
HAM	Mean	*0.019*	**0.050**	**0.033**	**0.056**
MSK	Max.	**0.066**	0.106	0.076	*0.068*
MSK	Median	**0.024**	0.026	**0.030**	*0.021*
MSK	Mean	**0.028**	0.035	0.031	*0.028*

Highest values for $\Delta^{max}\ S^{CD}_{b,d,m}$ per f^{scr} highlighted in **bold**, lowest values in *italics*

This difference between concepts c_{nevus} and $c_{melanoma}$ is furthermore observable in the other f^{tcn}. Models trained on f^{tcn}_{BCN} and f^{tcn}_{HAM} seem to learn and use c_{nevus} concepts significantly more than $c_{melanoma}$ concepts. Lastly there exists at least one m with S^{CD} close to 1 for $S^{CD}_{b_1,d_1}$ and $S^{CD}_{b_1,d_2}$. Thus, a selection of m by high concept knowledge and usage would have been possible.

5. CONCLUSIONS

This paper has examined the concept knowledge learned by different NN models and quantified using TCAV. To enhance comprehensibility and to unify the direction of the detected concept knowledge, concept detection scores CDS were introduced. Before evaluating the concept knowledge detectability of different models on different datasets, the validity of the results was established. The primary focus of the paper was the examination of concept knowledge detected from three different domain-shifted datasets and the assessment of the influence of these domain shifts on the detectable concept knowledge. The results showed that concept knowledge is largely domain-shift agnostic, meaning the models can apply their learned concept knowledge to new domains without suffering from generalizability issues that affect their predictive capabilities. Thus, for the tested use-case of interpretability for SKL, TCAV seems to be a viable interpretability tool which can be used without effort to manage generalizability (at least w.r.t interpretability). Since concept knowledge and usage seems to bridge domains, at least for the tested models, further research should be done towards the exploration of the relationship between concept knowledge and predictive capabilities within and across domains. If a positive relationship could be established, then NN model selection for high concept knowledge would be a way to address generalizability w.r.t. predictive capabilities. Lastly, investigating the generalizability of TCAV (and CDS) to applications beyond SKL, along with a comparison to other xAI methods would be a valuable research avenue.

REFERENCES

[1] A. J. London, "Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability," *The Hastings Center Report,* vol. 49, no. 1, p. 15–21, 2019.

[2] Y. Zhang, P. Tiňo, A. Leonardis and K. Tang, "A Survey on Neural Network Interpretability," *IEEE Trans. Emerg. Top. Comput. Intell.,* vol. 5, no. 5, p. 726–742, 2021.

[3] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas and R. Sayres, "Interpretability Beyond Feature Attribution: Quantitative Testing with Concept Activation Vectors (TCAV)," in *Conference: 35th International Conference on Machine Learning (ICML)*, Stockholm, Sweden, 2018.

[4] R. V. Zicari, S. Ahmed, J. Amann, S. A. Braun, J. Brodersen, F. Bruneault, J. Brusseau, E. Campano, M. Coffee, A. Dengel, B. Düdder, A. Gallucci, T. K. Gilbert and P. Gottfrois, "Co-Design of a Trustworthy AI System in Healthcare: Deep Learning Based Skin Lesion Classifier," *Front. Hum. Dyn.,* vol. 3, 2021.

[5] G. Argenziano, G. Fabbrocini, P. Carli, V. d. Giorgi, E. Sammarco and M. Delfino, "Epiluminescence microscopy for the diagnosis of doubtful melanocytic skin lesions. Comparison of the ABCD rule of dermatoscopy and a new 7-point checklist based on pattern analysis," *Archives of dermatology,* vol. 134, no. 12, p. 1563–1570, 1998.

[6] A. Lucieri, M. N. Bajwa, S. A. Braun, M. I. Malik, A. Dengel and S. Ahmed, "On Interpretability of Deep Learning based Skin Lesion Classifiers using Concept Activation Vectors," in *2020 International Joint Conference on Neural Networks (IJCNN)*, Glasgow, UK, 2020.

[7] A. Lucieri, M. N. Bajwa, S. A. Braun, M. I. Malik, A. Dengel and S. Ahmed, "ExAID: A multimodal explanation framework for computer-aided diagnosis of skin lesions.," *Computer methods and programs in biomedicine,* vol. 215, 2022.

[8] I. Goodfellow, Y. Bengio and A. Courville, Deep learning (Adaptive computation and machine learning), Cambridge: MIT Press, 2016.

[9] K. Fogelberg, S. Chamarthi, R. C. Maron, J. Niebling and T. J. Brinker, "Domain shifts in dermoscopic skin cancer datasets: Evaluation of essential limitations for clinical translation," *New biotechnology,* vol. 76, p. 106–117, 2023.

[10] J. Kawahara, S. Daneshvar, G. Argenziano and G. Hamarneh, "7-Point Checklist and Skin Lesion Classification using Multi-Task Multi-Modal Neural Nets," *IEEE journal of biomedical and health informatics,* vol. 23, no. 2, p. 538–546, 2018.

[11] N. Codella, V. Rotemberg, P. Tschandl, M. E. Celebi, S. Dusza, D. Gutman, B. Helba, A. Kalloo, K. Liopyris, M. Marchetti, H. Kittler and A. Halpern, "Skin Lesion Analysis Toward Melanoma Detection 2018: A Challenge Hosted by the International Skin Imaging Collaboration (ISIC)," in *2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018)*, Washington, D.C., USA, 2018.