Domain Specificity in Supervised Machine Learning Analogies

A Comparative Study of General Domain vs. Gaming Domain Analogies

Mateo Alexander Nasse

Supervisor(s): Gosia Migut, Ilinca Rențea, Yuri Noviello

EEMCS, Delft University of Technology, The Netherlands

A Thesis Submitted to EEMCS Faculty Delft University of Technology,
In Partial Fulfilment of the Requirements
For the Bachelor of Computer Science and Engineering
June 22, 2025

Name of the student: Mateo Alexander Nasse
Final project course: CSE3000 Research Project
Thesis committee: Gosia Migut, Ilinca Rențea, Yuri Noviello, David Tax

Abstract

This research paper looks into the influence of domain specificity on the understanding and motivation of first-year computer science students learning different concepts in supervised machine learning. Two types of domains were chosen for the analogies, the general domain and the gaming domain, the latter being the more specific one. These were evaluated in two phases. First, experts rated the analogies based on different metrics. Then, a user study was carried out using A/B testing to measure knowledge gain and motivation when exposed to the analogies. Results from the user evaluation show no statistically significant differences in terms of understanding for domain-specific or general analogies. Motivation, similarly show little difference when comparing both domains. The findings suggest that if analogies are helpful when it comes to understanding a topic, as long as the learner knows the domain, they do not play a big role.

1 Introduction

From personalized recommendations on streaming platforms to virtual assistants like Chat-GPT, machine learning (ML) powers many of the tools humans rely on every day. Yet, despite their widespread presence, the inner workings of these technologies are still a mystery to most users. This dependence raises questions regarding our control over systems of the future. For such reason, learning machine learning is not only important, it is also necessary in our increasingly algorithm-driven world.

However, the layered complexity of concepts, each building on top of the last, and the tendency to rely on mathematical explanations leave learners overwhelmed [1]. This creates a steep and discouraging entry point into a field that is becoming increasingly essential for a wide range of disciplines.

A promising approach to combat this challenge is the use of analogies in teaching, including analogies generated by language models (LM). The use of these analogies can help clarify ML concepts for both humans and even machines [2], and they are prevalent in emerging fields such as Explainable AI [3]. Yet, despite the important role of similarity in machine learning, few educational resources use analogies to teach these concepts to humans, and literature in this area remains limited [4; 5].

Formulating an analogy involves comparing a familiar 'source' concept (the analog) to a less familiar 'target' concept making use of shared features [6]. The domain chosen for the analogy tends to be general to expand the audience that can understand it, but the possibility of sacrificing the reach of the domain to target individuals with deep knowledge of it could lead to better understanding [7; 8].

Nevertheless, it remains unclear whether tailoring analogies to the learner's domain of expertise actually improves understanding. To investigate this, the present study presents analogy-based explanations for key concepts in supervised machine learning, a sub-set of ML known for its abstract and mathematically heavy concepts that tend to pose difficulty for novice learners [9]. Specifically, it will examine whether domain-specific analogies (in gaming) aimed towards individuals familiar with the domain lead to better understanding in comparison to general-domain analogies. This paper proposes the following questions to guide the study:

- Research Question: How does domain specificity influence analogy-based explanations in supervised machine learning for first-year bachelor computer science students?
- Sub-Question: How do analogy-based explanations using computer gameplay concepts influence knowledge gain and motivation in first-year computer science students learning supervised machine learning?
- Sub-Question: How do expert evaluations rate the quality of analogies with different specificity for teaching supervised machine learning, based on predefined criteria?

Section 2 gives background information and the rationale for choosing gaming as a domain. Section 3 discusses the topics from which analogies are generated and how they were generated. Section 4 talks about the methodology for the expert evaluation, the creation of the survey, and how that leads to a subset of analogies with which to proceed the study. Section 5 similarly to Section 4 starts with the methodology for the user evaluation and the creation of the A/B testing survey to carry out the evaluation with. Section 6 presents the results of the evaluation and an analysis of the meaning of the data. Section 7 outlines the ethical and responsible practices that were followed throughout the study. Section 8 discusses the findings and interprets them taking into consideration the research questions. Section 9 summarizes the main findings, restates the research questions and their answers. Finally, it provides recommendations for future work based on this study's results.

2 Background: Analogies in Education and Machine Learning

Analogies are crucial to how humans understand unfamiliar or complex concepts through their mapping onto familiar experiences. Hüllermeier [3] defines an analogy as a relationship between four objects where the way A relates to B is the same as the way C relates to D. Everyday examples, such as comparing the Sun to a light bulb or describing the expansion of the universe with a raisin bread, showcase the common use of analogies in communication and education [10]. In this context, mapping refers to the process of matching features between the pairs ($A \rightarrow B$ and $C \rightarrow D$) such that the structure of the relationship between A and B is preserved in the relationship between C and D. Gentner's [11] structure-mapping theory puts an emphasis on successful analogies that do not require deep knowledge of both the source and target domains. Rather, they depend on recognizing relationships in structure. This feature makes analogies most useful to novice learners, who benefit from a more abstract understanding when approaching unfamiliar topics [7].

In computer science education, analogy-based explanations are commonly used to teach abstract concepts in intuitive ways [12]. Although analogies tend to be used in areas such as data structures and algorithms, their use in ML education is more limited. This could be due to the perception that ML's abstract concepts of mathematical nature resist simplification through analogies [3]. However, multiple successful examples challenge this idea. For instance, the "going down a mountain" analogy for gradient descent is widely used to illustrate it. The use of this analogy goes to show students find the process of learning with this method more engaging [13]. That said, not all analogies are beneficial. When analogies are poorly constructed, it can lead to misconceptions, especially when the mapping between the analogy's source and target is weak or misleading [14]. For this method of teaching to be effective, it is important that the analogy is both understandable and correctly mapped.

One approach to making analogies more effective is to make sure the learner is familiar with the analogy's source domain. Traditionally, the source is drawn from a widely familiar domain such as nature, household activities, or eating, ensuring familiarity [15]. However, domain-specific analogies, which are drawn from more specialized areas, may bring more engagement and understanding for certain audiences. Even though such analogies might have a reduced reach, they could offer stronger conceptual understanding for learners in the target group. In the current study, domain-specific analogies are taken from the world of gaming. The rationale is that many computer science students are familiar with gaming mechanics, narratives, and systems [16]. Nonetheless, the risk of overfitting the analogies by focusing of sub-genres of games is still present. To mitigate this risk, the analogies in this study stick to recognizable gaming concepts such as game genres, player styles, and common elements in games.

This study focuses on applying the analogies to the teaching of supervised machine learning. By including these concepts into familiar gaming scenarios, the aim is to investigate whether the specificity of the analogies can be an effective way of learning. This approach offers insights on a different approach to traditional theoretical teaching when it comes to ML.

3 Design and Generation of Analogies

This study focuses on comparing the specificity of analogies and how effective are they in helping understand the concepts they explain. To achieve this, first, a list of commonly seen supervised machine learning concepts was taken to make analogies of. The reason to choose some of the common concepts is to make the analogies prevalent. These analogies are later used in the expert and user evaluations. The overall process followed in the study is illustrated in Figure 1. The concepts chosen are as follows:

- Curse of Dimensionality
- Discriminative vs Generative Classifiers
- Nearest Centroid Classifier
- K Nearest Neighbors Classifier
- Parzen Classifier
- Linear Discriminant Analysis
- Quadratic Discriminant Analysis
- Linear vs Non-linear
- Gaussian Naive Bayes
- Dimensionality Reduction

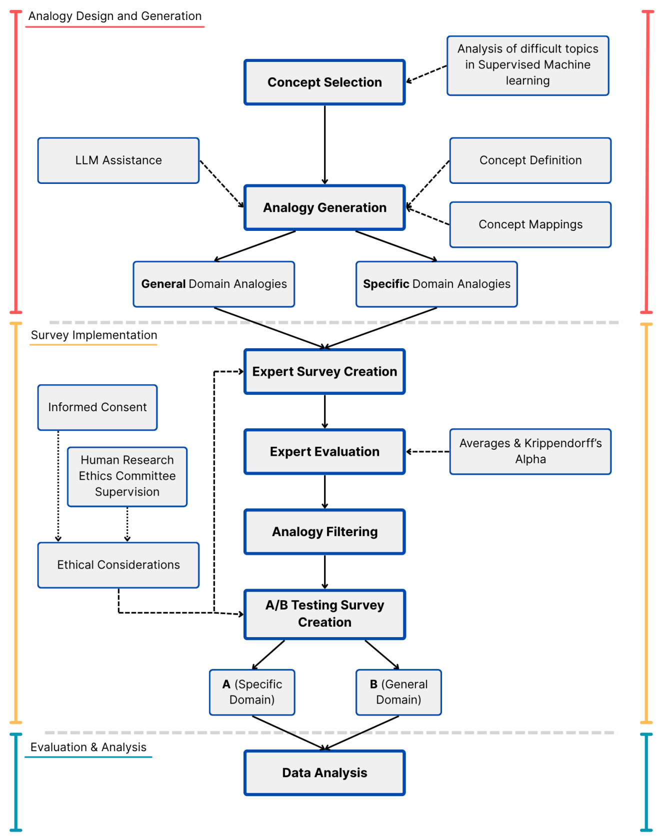

Figure 1: Overview of the methodology followed throughout the study.

Analogy Generation

The creation of analogies goes through a process to ensure they are sound. In this case, a concept diverged into two analogies. One of the analogies was part of the general domain, and the other part of the gaming domain. The process to create an analogy starts from the definition of the concept in question. It is important to note which are the features of the definition which should be mapped into features in the analogies. Furthermore, based on a study [17] and ongoing research[1], analogies can be measured and rated by target concept coverage, mapping strength, and metaphoricity. Target concept coverage refers to how well the concepts in the definition chosen for the topic are covered in the analogy. To make sure this was the case for all analogies, as mentioned before, the features of importance in the description

[1]Research can be found at: https://sites.google.com/illinois.edu/analogyeval24

were identified and listed. Furthermore, it was made sure that the connection between those features and those made up in the analogy was sound and consistent, meaning that the roles they have in each scenario are preserved. This is what mapping strength refers to. Lastly, the metaphoricity of an analogy is measured by the conceptual distance between the original concept and the target concept. For instance, explaining how a car moves compared to a bike does not have as much metaphoricity as comparing it to a pizza.

LLM Assistance

Coming up with the analogies was a process that used AI assistance, specifically ChatGPT suggestions with the GPT-4o model. Starting with the topics for the analogies, prompts were used to find elements that would correctly map the topics to enhance the criteria by which the analogies were measured. An example of the prompts can be seen in Appendix A. The use of ChatGPT to formulate analogy-based explanations and related information has increased in recent years as humans tend to ask for simpler and more intuitive ways of explaining concepts to understand them [18]. Initially, 20 analogies were created to be evaluated by experts; these can be seen in Appendix B, and after their evaluation, a subset of them was provided to users for further evaluation. This set of analogies can also be found on a website[2] where other researchers have uploaded similar analogies for machine learning.

4 Expert Evaluation Methodology and Analogy Selection

To assess analogy quality, a survey was given to experts containing a randomized ordering of the analogies from the analogy bank. Each analogy was rated on a 3-point Likert scale (1 = low, 3 = high) across three criteria adapted from Bhavya et al. [17]: target concept coverage, mapping strength, and metaphoricity. The survey was designed to take 10-20 minutes. Given the large amount of analogies, participants were not required to evaluate the full set, but rather, however many they could answer in the recommended time. This led to not all analogies being reviewed by the same number of experts. However, the sampling was balanced to ensure analogies were evaluated multiple times.

Expert Evaluation

The expert group consisted of a mixture of bachelor's and master's students, teaching assistants and professors, all with experience in the machine learning field. At a minimum, they had completed a university-level machine learning course, meaning a basic understanding of the concepts being evaluated was guaranteed. A total of 15 experts filled in the survey, with four of them having machine learning knowledge going past the bachelor course. The aim was to have each analogy evaluated by 3 to 4 experts to ensure sufficient feedback for an analysis. Research shows that this number of expert evaluations per concept is sufficient for a comprehensive evaluation [19]. Furthermore, to keep the validity of responses, experts were told to skip gaming-based analogies if they were not familiar with the domain. This made sure that domain-specific analogies were only judged by those who understood their context.

Selection Based on Expert Ratings

Given the large amount of contribution, most analogies were reviewed by at least three experts. Two metrics were measured, one being the average score (1 to 3 scale), and the other being Krippendorff's Alpha (KA) which allows us to measure inter-rater reliability [20]. Both general and domain-specific analogies needed to be well-formulated to proceed with the user evaluation. These were ranked according to the average score and KA to assess their overall quality. Based on this ranking, it was observed that two concepts (K-NN Classifier and Gaussian Naive Bayes Classifier) had well rated analogies in both the general and specific domain. These two were selected for the next phase of user testing due to scoring higher than most as well as having a similar rating across domains. See Table 1 for the full results. Even tough the Krippendorff's Alpha for most analogies is not high enough to say that the experts are in full agreement over the analogies, the selected analogies were still formulated well enough to be used in the user evaluation. The differences in agreement could show the subjectivity of interpreting analogies rather than the analogies themselves being flawed [21]. The analogies chosen are as follows:

K Nearest Neighbors Classifier

- **General Domain:** A doctor wants to diagnose a patient with a given set of symptoms. The doctor looks at a database for k patients with the most similar symptoms and diagnoses the most common diagnosis among them.
- **Gaming Domain:** To place a player in a matchmaking tier, it will look at the k nearest players with the most similar stats and put you in the most common tier among them.

Gaussian Naive Bayes

- **General Domain:** You want to recommend a movie to a friend, you know the movie genre, main actor/actress and studio. These influence your friends opinion independently, Based on other movies you know they have liked before you estimate how much they will like this one.
- **Gaming Domain:** You want to predict a Pokémon's type (Fire, Water, Electric, etc.) to gain an advantage in battle. You've observed the types of moves it uses, the kinds of opponents it tends to switch into, and its general battle behavior. Each of these clues influences your guess independently and equally. Based on similar Pokémon you've seen before, you estimate the most likely type for this one.

5 User Evaluation Methodology and A/B Survey Design

The target group for this study is first-year bachelor's in computer science. The goal sample size for this evaluation was a minimum of 20 students. There are two main reasons why

[2] The analogies are part of https://ml-teaching-analogies.github.io/

Concept	Avg_{Spec}	Avg_{Gen}	KA_{Spec}	KA_{Gen}
Gaussian Naive Bayes Classifier	3.00 ± 0.00	2.60 ± 0.63	1.000	0.452
K-NN Classifier	2.58 ± 0.67	2.33 ± 0.71	-0.106	-0.129
Discriminative vs Generative	2.33 ± 1.03	2.29 ± 0.78	-0.250	-0.087
Linear vs Nonlinear	2.25 ± 0.61	2.33 ± 0.78	-0.096	0.328
Dimensionality Reduction	2.25 ± 0.75	2.00 ± 0.71	0.441	-0.111
Nearest Centroid Classifier	2.72 ± 0.46	2.11 ± 0.60	-0.090	-0.233
Quadratic Discriminant	2.11 ± 0.90	2.17 ± 0.72	0.046	0.043
Parzen Classifier	2.00 ± 0.74	2.47 ± 0.64	0.083	-0.141
Linear Discriminant	1.86 ± 0.92	2.33 ± 0.78	0.146	-0.146
Curse of Dimensionality	1.44 ± 0.53	2.00 ± 0.87	0.600	-0.185

Table 1: Mean (AVG) $\pm$ standard deviation and Krippendorff's Alpha (KA) for Specific (Spec) and General (Gen) analogies.

computer science students were chosen; the first has already been mentioned in Section 2 stating the tendency of computer science students to know about the gaming domain given the trend of people familiar with that domain to choose the career. The other reason is that for most academic curricula in computer science, machine learning will be an area of study, and so, students are at least familiar with previous concepts required to understand it, such as statistics. The target group had to be first-years due to them not yet taking the machine learning course, and thus reduce the possibility of those taking the survey already having prior knowledge on the topics discussed. Following the expert evaluation, a user study with a selection of the best rated analogies was conducted to test the effectiveness of domain-specific analogies in comparison to those in the general domain. To facilitate this process, the users were evaluated by the "understanding" level in Bloom's taxonomy [22]. The questions for the user evaluation survey were created with this in mind. Additionally, the evaluation followed an A/B testing design using two groups of participants. The first group received only domain-specific analogies, while another group received general domain analogies.

User Evaluation Survey Creation

The user evaluation survey was designed to test difference in specificity. For this reason, two versions of the survey were created. The first included the general domain analogies for Gaussian Naive Bayes Classifier and K-NN Classifier, and the second contained the domain specific analogies. The survey started with the definition of the concept from which the analogies are based on. It continued by showing the analogy for the concept and three multiple choice questions regarding the concept. Across both versions of the survey the questions were generally consistent and only replaced the context of the analogy. Similarly, the possible answers were the same, with different wording to be understood in the context of the domain. This way, the possibility of bias by making questions in one survey easier can be reduced. Furthermore, every question contained the option "I'm not sure" which is encouraged to be checked in the case the user did not understand or is not sure of the answer. Adding this option should prevent guesswork and improve data accuracy [23]. The questions can be found in Appendix C. Through A/B testing, both surveys were distributed to collect a similar amount of responses. Given that not everyone is familiar with the gaming domain, which one survey was based on, the process was to first ask about the knowledge of this domain and try acquire the planned number of responses before the general domain survey. Both surveys ended with a reduced version of the Instructional Materials Motivation Survey (RIMMS) [24] that serves as a way to quantitatively measure a learner's motivation towards analogies in this case. The survey measured attention, confidence, relevance, and satisfaction based on short Likert scale statements about the material.

6 Results

A total of 24 first-year bachelors filled in the survey. The survey with gaming-related analogies was answered ten times, whilst the general analogies was answered fourteen times. For all questions related to the understanding of the topic, the percentage of right answers can be calculated. This data can give an insight on how much knowledge is gained through exposure to the definition and the analogies with different specificity. Afterwards, thanks to the Reduced Instructional Motivation Materials Survey (RIMMS), the motivation towards this instructional method can be measured.

Knowledge Gain

To measure knowledge gain, there is a need to determine if the accuracy between participants who took the general domain survey versus the domain-specific survey is statistically significant. For this research, Welch's t-test is used. Since Welch's test does not assume equal variances and is more reliable for unequal and smaller sample sizes, it is preferred to Student's t-test [25]. This test was performed on the user accuracy scores, what percentage of questions an individual participant has gotten correct, for all participants in both groups. The test revealed no statistical significance between the groups. The resulting t-statistic was -0.0715, a value that is close to zero, which indicates little to no difference between the two group means. Additionally, the corresponding p-value was 0.9439, meaning that there exists a 94.39% probability that the the difference observed could occur on random chance. The standard significance threshold is 0.05, given that the results are far above this threshold we can assume no statistical significance. For the general domain group, the average correctness was 44.05% while it was 43.33% for the domain specific group, as can be seen in Figure 2. These findings suggest that, at least for this sample, the domain chosen for the analogy does not have measurable effect on the participants knowledge gain as reflected by their accuracy answering the questions.

User Motivation

User motivation was assessed using the RIMMS, which measures the following four dimensions:

- **Attention**: Participants' perceived interest in the material.
- **Relevance**: The extent to which the analogies related to their background.
- **Confidence**: How confident participants felt in their understanding.

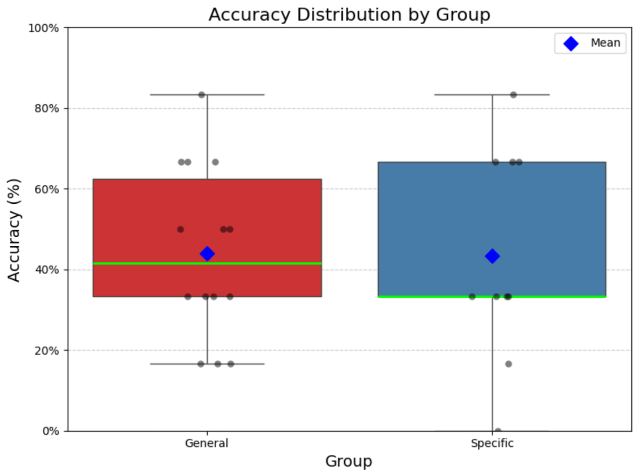

Figure 2: Accuracy scores by group. Boxplots show medians and IQRs, black dots are individual scores, and blue diamonds mark means.

- **Satisfaction**: Overall satisfaction with the explanation style.

Responses were recorded on a five-point Likert scale, where 1 means "not true" and 5 means "very true." Table 2 presents the average scores across dimensions for both groups of participants.

Dimension	Avg Score General	Avg Score Specific
Attention	3.4 ± 0.96	3.0 ± 1.16
Relevance	3.3 ± 1.04	3.2 ± 1.37
Confidence	2.8 ± 1.25	2.8 ± 1.13
Satisfaction	3.3 ± 0.81	3.2 ± 1.29

Table 2: Comparison of Average Scores $\pm$ standard deviation of Different Surveys Across Dimensions

Overall, the scores indicate that participants moderately agree that analogies, when paired with the definition of a concept, enhanced their motivation. However, the difference in motivational impact between general domain and domain specific analogies is minimal in all four dimensions. This suggests that even though analogies can positively contribute to learning engagement, the choice of domain does not significantly influence that motivation perceived.

7 Responsible Research

Use of Language Models in Analogy Creation

Some analogies used in this study were partially assisted by OpenAI's GPT-4o model. The large language model (LLM) was used to make sure analogies were strongly mapped through multiple features. Afterwards, all content was reviewed and validated. For grammar and wording, Writefull was used.

Data Collection

The study follows ethical standards for research with human participants. Ethical approval was given by the Human Research Ethics Committee (HREC). Furthemore, all participants were provided information about the study and the effect of their involvement and informed consent was obtained before participation. No personally identifiable information was gathered. Responses were stored securely and only used for the purposes of this research.

Bias Awareness

Possible bias was acknowledged during the study. To mitigate risks, randomized A / B testing was used, a wide set of analogies, and expert feedback from multiple evaluators. The subjectivity of interpretation was also taken into account as a limitation.

Replicability

For the purposes of replicability, the study provides a transparent and clear description of the methodology, including how the surveys were designed, the number and general profile of participants , and the statistical analyses performed to give meaning to the data. All analysis was conducted using public tools such as Microsoft Forms, Python and scientific libraries (e.g, SciPy, NumPy) to ensure the analysis is transparent and possible to replicate. Furthermore, appendices, containing survey questions, analogies and LLM prompts have been included.

8 Discussion

Expert Evaluation Findings

Experts assessed the analogies using three criteria: target concept coverage, mapping strength, and metaphoricity. Analogies from both domains showed a wide range of scores. Interestingly, the concept being explained tended to receive consistent ratings across the general and gaming domains, suggesting that clarity and relevance of the analogy may matter more than its domain origin. Nonetheless, inter-rater reliability was variable, a result that could be explained by both the subjective nature of evaluating analogies and the small number of evaluators per analogy.

User Evaluation: Knowledge Gain

A/B testing with students compared accuracy between the two domains in terms of understanding. Results revealed no statistically significant difference in correctness between groups. Welch's t-test gave $t = -0.0715$, $p = 0.9439$ indicating neither group outperforming the other. Even though specific analogies led to better accuracy in some questions, they performed worse in others. This variability can mean that domain familiarity may help learning in selective question types but it does not impact overall knowledge gain. This opens a possible opportunity for further research, especially in identifying when domain specificity can be helpful.

User Evaluation: Motivation

The RIMMS measured attention, relevance, confidence, and satisfaction on a five-point Likert scale. The averages across both groups were moderate (~ 3), with little difference between groups exposed to different domains. This suggests that while moderately motivating, analogies with specific domains do not influence motivation outcome any more than those in the general domain. One explanation could be students being equally familiar with both domains, meaning students who were given gaming analogies were not familiar with the domain to an advanced degree compared to their familiarity with the general domain. Future studies should make sure participants' prior knowledge of the specific domains is deeper.

Synthesis

Overall, domain specificity in analogies had limited impact on learning outcome and motivation. Most importantly, it did not hinder understanding, implying that good analogies from any familiar domain can be an effective instructional tool. The study also strengthens the idea that analogies, regardless of their specificity, can make the process of learning complex machine learning topics more simple, when matched to the learners' background. While previous work [7] has shown that analogies can enhance learning when learners are familiar with the source domain, this paper goes further into this understanding by comparing how the specificity of those analogies has an effect. It was found that whether an analogy came from the general or the gaming domain, there was no significant difference in knowledge gain or motivation. This might highlight the importance of familiarity rather than specificity.

Limitations

Several limitations can be identified in the study. The small sample size (24 participants split into groups of 10 and 14) reduces generalizability. Furthermore, participants' familiarity with the specific domain may not have been high enough to get a noticeable result. In the expert evaluation, some analogies were evaluated by different numbers of experts, which could have affected reliability. Moreover, the absence of qualitative feedback also limits understanding of participants' perceptions.

9 Conclusions and Future Work

This study explored how domain specificity influences analogies to explain concepts in supervised machine learning for first-year bachelor computer science students. It focused on three core research questions:

- Research Question: How does domain specificity influence analogy-based explanations in supervised machine learning for first-year bachelor computer science students?
- Sub-Question: How do analogy-based explanations using computer gameplay concepts influence knowledge gain and motivation in first-year computer science students learning supervised machine learning?
- Sub-Question: How do expert evaluations rate the quality of analogies with different specificity for teaching supervised machine learning, based on predefined criteria?

Findings suggest that while analogies can be effective as an instructional tool, domain specificity alone does not significantly influence a learners understanding or motivation. Expert evaluations showed both general and gaming domain analogies received similar scores in target concept coverage, mapping strength, and metaphoricity, with some variation which could be due to the subjectivity of evaluation and the limited number of raters per analogy. On a similar note, the user evaluation revealed no statistically significant differences between groups when it comes to knowledge gain and motivational scores measured though the Reduced Instructional Materials Motivation Survey, which was moderate for both groups. The results imply that for analogies to be effective, there is no requirement for a degree of specificity, as long as the domain is known to the reader. However, this study does come with limitations: the small sample size and uneven group distribution restrict how much understanding we can have on the subject. Future work should try to include a larger sample and incorporate qualitative data, and explore analogies based on a broader domains that are still semi-specific such as sports or pop culture. Furthermore, the effect of domains on the types of questions can also be studied. Ultimately, this research strengthens the idea that even if varying levels of domain familiarity play a role in certain contexts, what matters most is having a well-structured analogy to support learning in complex technical domains such as machine learning.

References

[1] Kayur Patel. Lowering the barrier to applying machine learning. In *CHI '10 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '10, page 2907–2910, New York, NY, USA, 2010. Association for Computing Machinery.

[2] Siyu Yuan, Cheng Jiayang, Lin Qiu, and Deqing Yang. Boosting scientific concepts understanding: Can analogy from teacher models empower student models? In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6026–6036, Miami, Florida, USA, November 2024. Association for Computational Linguistics.

[3] Eyke Hüllermeier. Towards analogy-based explanations in machine learning. In Vicenç Torra, Yasuo Narukawa, Jordi Nin, and Núria Agell, editors, *Modeling Decisions for Artificial Intelligence*, pages 205–217, Cham, 2020. Springer International Publishing.

[4] Vishnu S. Pendyala. Relating machine learning to the real-world: Analogies to enhance learning comprehension. In Kanubhai K. Patel, Gayatri Doctor, Atul Patel, and Pawan Lingras, editors, *Soft Computing and its Engineering Applications*, pages 127–139, Cham, 2022. Springer International Publishing.

[5] Ignacio Evangelista, Germán Blesio, and Emanuel Benatti. Why are we not teaching machine learning at high school? a proposal. In *2018 World Engineering*

Education Forum - Global Engineering Deans Council (WEEF-GEDC), pages 1–6, 2018.

[6] R A Sukamto, H W Prabawa, and S Kurniawati. Analogy mapping development for learning programming. *Journal of Physics: Conference Series*, 812(1):012109, feb 2017.

[7] Jason L. G. Braasch and Susan R. Goldman. The role of prior knowledge in learning from analogies in science texts. *Discourse Processes*, 47(6):447–479, August 2010.

[8] Bianca A. Simonsmeier, Maja Flaig, Anne Deiglmayr, Lennart Schalk, and Michael Schneider. Domain-specific prior knowledge and learning: A meta-analysis. *Educational Psychologist*, 57(1):31–54, January 2022.

[9] Falco J. Bargagli Stoffi, Gustavo Cevolani, and Giorgio Gnecco. Simple models in complex worlds: Occam's razor and statistical learning theory. *Minds and Machines*, 32(1):13 – 42, 2022. Cited by: 34.

[10] Puripant Ruchikachorn and Klaus Mueller. Learning visualizations by analogy: Promoting visual literacy through visualization morphing. *IEEE Transactions on Visualization and Computer Graphics*, 21(9):1028–1044, September 2015.

[11] Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155 – 170, 1983. Cited by: 3505; All Open Access, Bronze Open Access.

[12] Pawan Saxena, Sanjay Kumar Singh, and Gopal Gupta. Achieving Effective Learning Outcomes through the Use of Analogies in Teaching Computer Science. *Mathematics*, 11(15):1–18, July 2023.

[13] Shang Shanshan, Du Chenhui, and Luo Lijuan. Metaphor and board writing matter: The mediating roles of psychological distance and immersion in video lectures. *Computers & Education*, 191:104630, 2022.

[14] Rand Spiro, Paul J. Feltovich, Richard Coulson, and Daniel Anderson. *Multiple analogies for complex concepts: Antidotes for analogy-induced misconceptionin advanced knowledge acquisition*, pages 498–530. Cambridge University Press, 01 1989.

[15] Reinders Duit. The role of analogies and metaphors in learning science. *Science Education*, 75:649 – 672, 11 1991.

[16] Zohal Shah, Chen Chen, Gerhard Sonnert, and Philip M. Sadler. The influences of computer gameplay and social media use on computer science identity and computer science career interests. *Telematics and Informatics Reports*, 9:100040, 2023.

[17] Bhavya Bhavya, Chris Palaguachi, Yang Zhou, Suma Bhat, and ChengXiang Zhai. Long-form analogy evaluation challenge. In Simon Mille and Miruna-Adriana Clinciu, editors, *Proceedings of the 17th International Natural Language Generation Conference: Generation Challenges*, pages 1–16, Tokyo, Japan, September 2024. Association for Computational Linguistics.

[18] Malik Sallam, Walaa Elsayed, Mona Al-Shorbagy, Mariam Barakat, Sally El Khatib, Wissam Ghach, Nisreen Alwan, Souheil Hallit, and Dana Malaeb. Chatgpt usage and attitudes are driven by perceptions of usefulness, ease of use, risks, and psycho-social impact: a study among university students in the uae. *Frontiers in Education*, 9:1414758, 2024.

[19] Jakob Nielsen. How to conduct a heuristic evaluation. *Nielsen Norman Group*, 1995.

[20] Klaus Krippendorff. Computing krippendorff's alpha-reliability. https://repository.upenn.edu/cgi/viewcontent.cgi?article=1043&context=asc_papers, 2011. University of Pennsylvania ScholarlyCommons.

[21] Gaole He, Agathe Balayn, Stefan Buijsman, Jie Yang, and Ujwal Gadiraju. It is like finding a polar bear in the savannah! concept-level ai explanations with analogical inference from commonsense knowledge, 2025. mailto: g.he@tudelft.nl.

[22] Lorin W. Anderson and David R. Krathwohl. *A Taxonomy for Learning, Teaching, and Assessing: A Revision of Bloom's Taxonomy of Educational Objectives*. Longman, New York, 2001.

[23] Marc T. Kiviniemi, Erin M. Ellis, Heather Orom, Erika A. Waters, and Jennifer L. Hay and. 'don't know' responding and estimates of perceived risk: failing to provide a 'don't know' response systematically biases laypeople's perceived risk estimates. *Health, Risk & Society*, 22(1):69–85, 2020.

[24] Nicole Loorbach, Oscar Peters, Joyce Karreman, and Michaël Steehouder. Validation of the instructional materials motivation survey (imms) in a self-directed instructional setting aimed at working with technology. *British Journal of Educational Technology*, 46, 03 2014.

[25] Tetsuya Sakai. Two sample t-tests for ir evaluation: Student or welch? SIGIR '16, page 1045–1048, New York, NY, USA, 2016. Association for Computing Machinery.

.

A ChatGPT Analogy Mapping Transcipt

The following interaction was done with different concepts but the same phrasing leading to similar results for other analogies.

Prompt

> I have two analogies for K-NN neighbors classifier. The analogies are in different domains and are both based on the following definition: When you want to classify a data point into a category (like spam or not spam), the K-NN algorithm looks at the K closest points in the dataset. These closest points are called neighbors. The algorithm then looks at which category the neighbors belong to and picks the one that appears the most. This is called majority voting. The general domain analogy is to compare the points in the data set to a set of symptoms a doctor is diagnosing in a patient. The doctor looks at the k patients with similar symptoms in a database and then makes a diagnosis. The specific domain analogy is to compare the points in the data set to the stats on other players when trying to assign someone a tier in a matchmaking system. The system would look at the k players with most similar stats and make a decision.
>
> Help me figure out everything that can be mapped between the definition and the analogies.

Response

ChatGPT response, formatted to fit the research paper.

Definition of k-NN

The k-Nearest Neighbors (k-NN) algorithm is a non-parametric method used for classification. To classify a new data point, the algorithm finds the k most similar data points in the dataset based on a distance metric. It then uses majority voting among these k neighbors to assign a class label to the new point.

Key Concepts in the k-NN Algorithm

Concept	**Meaning in Machine Learning**
Data point (to classify)	A new, unlabeled example
Dataset	Set of labeled data points with known categories (e.g., spam vs. not spam)
k nearest neighbors	The k most similar data points (according to a distance metric)
Distance/similarity	Measure of closeness or similarity between points
Category (class label)	The known class (e.g., spam, not spam; gold tier, platinum tier)
Majority voting	The algorithm selects the most common class among the k neighbors
Final classification	The predicted class assigned to the new data point

Table 3: Key elements in k-NN classification

General Domain Analogy: Medical Diagnosis

A doctor is diagnosing a patient. They look at the k patients with the most similar symptoms in their medical database and choose the diagnosis that appears most frequently among them.

k-NN Concept	**Medical Analogy Element**
Data point (to classify)	Current patient with unknown illness
Dataset	Medical database of past patients
k nearest neighbors	k most similar past patients
Similarity metric	Symptom similarity (e.g., fever, cough, pain)
Category (class label)	Diagnosis (e.g., flu, cold, pneumonia)
Majority voting	Most common diagnosis among similar cases
Final classification	Diagnosis assigned to the current patient

Table 4: Mapping between k-NN and medical diagnosis analogy

Gaming Domain Analogy: Matchmaking System

A matchmaking system is assigning a player to a skill tier. It looks at the k most similar players based on gameplay statistics and assigns the new player the tier that is most common among those similar players.

k-NN Concept	**Gaming Analogy Element**
Data point (to classify)	New player with unknown skill tier
Dataset	Database of players with known stats and tiers
k nearest neighbors	k most similar players based on stats
Similarity metric	Performance similarity (e.g., win rate, K/D ratio, ELO)
Category (class label)	Skill tier (e.g., Bronze, Silver, Gold)
Majority voting	Most frequent tier among k similar players
Final classification	Tier assigned to the new player

Table 5: Mapping between k-NN and matchmaking system analogy

Shared Insights Across Analogies

- **Similarity as distance:** Both analogies rely on closeness — whether by symptoms or gameplay stats — to define neighborhood.
- **No training phase:** k-NN does not build a model; it simply stores data and queries it at classification time.
- **Sensitivity to k:** A small k can lead to overfitting or noise; a large k might dilute precision.
- **Local decision-making:** The classification is made using only nearby (similar) examples, not the entire dataset.

Bonus Insight

These analogies illustrate why k-NN is both intuitive and computationally expensive. Doctors and matchmaking systems both rely on historical data and make decisions based on real-time comparisons to similar cases. These metaphors also highlight potential failure cases when the data is sparse, noisy, or high-dimensional—providing a conceptual bridge to topics like the *curse of dimensionality* and *feature selection*.

B List of Analogies

Curse of Dimensionality

- **General Domain:** The curse of dimensionality is like having to find a lost sock in your house. The more rooms in the house, the harder it will be to find.
- **Gaming Domain:** The curse of dimensionality is like trying to learn combo moves in a fighting game. When the combo requires too many inputs, it becomes harder to execute.

Discriminative vs Generative Classifiers

- **General Domain:** Imagine you have to guess someone's profession by looking at them.
  - Discriminative approach: You know that anyone who has worn a tan suit works in marketing while someone with a black suit works in legal.
  - Generative approach: You try to think how someone who works in marketing dresses, acts, or walks. You check if your vision of a marketing employee matches the employee you see.
- **Gaming Domain:** You classify players' behaviour.
  - Discriminative approach: You know that if a players tag starts with "xX" and ends with "Xx" they are a camper.
  - Generative approach: You know all about players' behaviours. You know what guns they use, how they move and where they tend to be. If a player keeps holding a corner with specific weapons you know they are likely a camper.

Nearest Centroid Classifier

- **General Domain:** You are at a party and there are multiple groups of people. You don't know where to hangout so you measure the average vibe per group and go to the one that feels closest to you.
- **Gaming Domain:** You are playing a strategy game with multiple factions that want to control territory and see a village appear at a point in the map. You guess which faction that village belongs to by picking the faction whose average building location is closest to the village.

K Nearest Neighbours Classifier

- **General Domain:** A doctor wants to diagnose a patient with a given set of symptoms. The doctor looks at a database for k patients with the most similar symptoms and diagnoses the most common diagnosis among them.
- **Gaming Domain:** To place a player in a matchmaking tier, it will look at the k nearest players with the most similar stats and put you in the most common tier among them.

Parzen Classifier

- **General Domain:** You're in a mall surrounded by many overlapping Wi-Fi signals. Each store has hotspots. Your phone estimates your location based on the signal strength from each brand's hotspots. The brand which has the strongest combined signal is likely where you are.
- **Gaming Domain:** In an open-world RPG, the music changes depending on what region you're in — desert, swamp, forest, etc. But instead of hard boundaries, the game blends music gradually based on how close you are to the center of each region. If you're standing near the edge of a swamp and a forest, you'll hear a blend of both, with louder swamp music if you're closer to the swamp center. The game computes the combined influence of all region centers near you, and plays the music from the region with the strongest combined signal. .

Linear Discriminant Analysis (LDA)

- **General Domain:** Imagine a university with Computer Science, Mechanical Engineering and Aerospace Engineering students. You want to send each student to their correct class but can't interview everyone. Linear Discriminant Analysis makes a rule (line) that separates students based on features such as their average grade on Machine learning and Linear Algebra.
- **Gaming Domain:** Imagine you are playing an action RPG game where you are fighting a dangerous boss. You want to read the boss' behavior to predict when to dodge. Linear discriminant analysis would be a rule that you follow that leads you to decide to dodge early or late. For instance you can choose to divide the two types of dodges based on whether the boss' eyes are looking left or right and the wind-up time of the attack.

Quadratic Discriminant Analysis (QDA)

- **General Domain:** You want to classify athletes into "low potential", "medium potential", and "high potential" based on their speed and endurance. It is easy to detect when someone has low potential but for a high potential there is way more variance. Quadratic Discriminant Analysis is a way to model each athlete when a simple rule like a straight line is not enough to differentiate them. item **Gaming Domain:** Imagine you are playing an action RPG game and want to estimate an enemy's difficulty before fighting them. You can see their size, calmness, and complexity in design. If you want to separate them into "easy", "medium", and "hard" to defeat, you employ rules to separate them. You notice it's easy to label when an enemy will be easy but it gets complicated to know when they are hard as there is a lot of variance in those enemies. Quadratic Discriminant Analysis is a way to model how difficult the enemy will be by applying curved boundaries to separate the groups.

Linear vs Non-linear Models

- **General Domain:** If you are on a road trip you can estimate the time it will take to finish the trip if more distance is introduced. If the distance and the time it took was linear, doubling the distance needed to travel would mean it would take double the time to finish the trip. This could happen if for instance the trip was one long highway. If the distance and the time it took was non-linear, doubling the distance needed to travel would not mean the time it would take was double. There are other factors that come into play such as big cities and their traffic.
- **Gaming Domain:** You are designing a skill tree for a game. There are two ways you could do this tree. The linear approach would be to grant a set amount of points per upgrade (eg, +5 damage per upgrade). The non-linear approach might be more useful to balance it as it would let you give small bonuses at first and then skyrocket into certain branches after investing a number of points into them.

Gaussian Naive Bayes

- **General Domain:** You want to recommend a movie to a friend, you know the movie genre, main actor/actress and studio. These influence your friends opinion independently, Based on other movies you know they have liked before you estimate how much they will like this one.
- **Gaming Domain:** You want to predict a Pokémon's type (Fire, Water, Electric, etc.) to gain an advantage in battle. You've observed the types of moves it uses, the kinds of opponents it tends to switch into, and its general battle behavior. Each of these clues influences your guess independently and equally. Based on similar Pokémon you've seen before, you estimate the most likely type for this one.

Dimensionality Reduction

- **General Domain:** Imagine you are on a roadtrip and you have a map of the roads and cities you are going through. The map contains every single street in the city and roads that you are not gonna go through. In order to focus on your path you tear off the parts of the map that are not useful to you.

- **Gaming Domain:** In open-world action adventure games, the use of the minimap is crucial to navigate the terrain. When you plan to travel to a keypoint, some minimaps allow you to hide irrelevant information so you can focus on the roads, terrain and your marker. By removing enemy locations, points of interest or shop icons the map is reducing the information into what is useful for you.

C User Evaluation Survey Questions

K-NN Questions

Q1: Interpreting Distance

General Analogy: In the doctor analogy, what does "distance" represent in terms of K-NN?

- A. The count of shared symptoms between patients
- B. How common a symptom is in the population
- **C. The degree of similarity between symptom profiles**
- D. The patient's likelihood of recovery

Gaming Analogy: In the matchmaking analogy, what does "distance" represent in terms of K-NN?

- A. The difference in rank history
- B. Whether players have competed before
- **C. How similar the players are based on gameplay stats**
- D. The average time spent playing each match

Q2: Curse of Dimensionality

General Analogy: Imagine the doctor now has access to a very large dataset of past patients with thousands of symptoms recorded. What challenge might this pose to the K-NN diagnosis approach?

- A. Too many symptoms might reduce noise
- **B. Similarity comparisons may become less meaningful**
- C. More data will improve generalization
- D. It allows for better disease ranking

Gaming Analogy: Imagine the matchmaking system tracks thousands of different stats per player. What challenge might this pose to placing players using K-NN?

- A. Rankings will be more consistent
- **B. Statistical similarity might become harder to measure**
- C. Matchmaking will be faster with more attributes
- D. Players will be split into narrower subgroups

Q3: Choosing the Right k

General Analogy: If the doctor uses too many past cases (a very large k) to make a diagnosis, what could go wrong?

- A. The model will converge to an average diagnosis
- **B. Irrelevant cases may dominate the prediction**
- C. Diagnostic accuracy will increase
- D. Outliers will be removed automatically

Gaming Analogy: If the system looks at too many players to decide your tier, what might go wrong?

- A. The player will always be placed in the median tier
- **B. Less relevant players may influence the outcome**
- C. Tier assignment becomes deterministic
- D. Rankings will always be more stable

Gaussian Naive Bayes Questions

Q1: Feature Weighting and Assumptions

General Analogy: In the movie recommendation analogy, suppose your friend really cares about genre but barely notices the studio. How does Gaussian Naive Bayes treat this?

- A. It learns which features are most predictive over time
- **B. It treats genre and studio as equally important**
- C. It removes features that add little information
- D. It ranks different genres above the studio

Gaming Analogy: In the Pokémon analogy, suppose the type of moves used by the opponent is very informative, but switching behavior adds little information. How does Gaussian Naive Bayes treat this?

- A. It gives more weight to move types than switches
- **B. It treats both clues as equally useful**
- C. It selects only the most predictive feature
- D. It balances features based on prediction accuracy

Q2: Independence Assumption and Performance

General Analogy: Why might a Naive Bayes classifier still perform well at recommending movies even if genre and actor preferences are slightly related?

- A. It uses genre–actor co-occurrence to adjust predictions
- **B. It still generalizes well despite the independence assumption**
- C. It retrains based on correlated feedback
- D. It reweights correlated inputs automatically

Gaming Analogy: Why might the classifier still guess a Pokémon's type well, even if move choice and switching behavior are somewhat related?

- A. It explicitly models the interaction between clues
- **B. Its predictions remain strong despite feature dependencies**
- C. It corrects itself using feedback from battles
- D. It relies more on the dominant clue

Q3: Robustness Despite Violations

General Analogy: Even if genre and actor aren't totally independent, why might this simple approach still work for predicting your friend's movie taste?

- A. It uses cross-validation to optimize performance
- B. It blends predictions from other classifiers
- **C. It can still make useful guesses even if assumptions aren't perfect**
- D. It learns a full joint distribution over all features

Gaming Analogy: Even if move types and switching behavior aren't independent, why might this model still be useful in battle?

- A. It memorizes all prior outcomes and counters
- B. It calculates exact conditional probabilities
- **C. It works well even when the assumptions don't fully hold**
- D. It updates its belief mid-battle