Prospects of Explainability in the Hugging Face Hub Landscape

Saša Brdnik*, Tjaša Heričko

Faculty of Electrical Engineering and Computer Science, University of Maribor, Koroška cesta 46, Maribor, Slovenia

Abstract

Machine learning models are widely adopted in intelligent systems to support decision-making across various domains, such as healthcare, finance, and law enforcement. Despite their increasingly remarkable performance, the opacity of these models poses substantial challenges, particularly in the aspect of explainability. The Hugging Face Hub has become a pivotal repository for publicly available machine learning models, extensively reused in both academia and industry. This paper aims to address the imperative for explainability by analyzing the current state of inherent and non-augmented perceived explainability of machine learning models shared on the platform. By data mining the model registry and exploring the models' documentation, the models were categorized based on five common classes of machine learning algorithms discussed and empirically investigated from end-users' perspectives regarding explainability in existing works. Building on theoretical and user-centred empirical evidence from prior works focused on model performance and explainability, this work attempts to contextualize these two dimensions within the real-world distribution and popularity of models on the hub, thereby introducing a third dimension. Additionally, this work examines explainability through a three-tier categorization of models based on their explainability, i.e., non-, mediocrely-, and highly-explainable. The findings highlight the disproportionate prevalence of models built on deep neural networks, which are considered among the least explainable compared to those based on other classes of learning algorithms.

Keywords

Hugging Face, model registry, machine learning, explainable artificial intelligence, XAI, explainability

1. Introduction

In recent years, the rapidly increasing use of advanced machine learning (ML) models has significantly enhanced the capabilities of various intelligent system applications to support decision-making across multiple domains. Following the increased demand for accessible ML models and to facilitate their reuse, the Hugging Face (HF) Hub has emerged as a pivotal repository of ML models [1, 2]. The models shared via the platform are widely adopted and reused across academia and industry due to their versatility and performance across a variety of tasks. Despite their remarkable performance, the opacity of these models poses substantial challenges, particularly in the realm of explainability [3]. Explainability in artificial intelligence (AI) is critical for ensuring transparency, trust, privacy awareness, and ethical AI deployment [4]. It is not solely a technical concern; it is a multifaceted issue extending to legal, ethical, and practical domains. For models to be reliable in sensitive applications such as healthcare, education, employment, finance, and law enforcement, stakeholders must understand *how* and *why* these models make specific decisions, recommendations, or actions [5]. Hence the recent interest in explainable artificial intelligence (XAI). This necessity has also been highlighted in a recent wave of legislative changes – the AI Act in the European Union [6] and a blueprint for the AI Bill of Rights in the USA [7]. Both emphasize the importance of explainability and developing techniques to explain the inner workings of ML models, such as those shared via the HF Hub.

Existing work in XAI has explored the explainability of models based on various ML algorithms [3, 4, 5, 8, 9]. However, this has not been contextualized in terms of the distribution and popularity of publicly shared ML models on leading platforms, such as the HF Hub. This work addresses the imperative for

SQAMIA 2024: Workshop on Software Quality, Analysis, Monitoring, Improvement, and Applications, September 9–11, 2024, Novi Sad, Serbia

*Corresponding author.

sasa.brdnik@um.si (S. Brdnik); tjasa.hericko@um.si (T. Heričko)

0000-0003-3730-2769 (S. Brdnik); 0000-0002-0410-7724 (T. Heričko)

explainability by analyzing the current state of end-user perceived explainability of ML models from the HF Hub. To achieve this, all models shared via the HF Hub at the time of conducting the research, which included more than 685k models, were obtained. Data mining was performed on the documentation related to the models to categorize them into five common classes of ML algorithms [5]. While the research community [10, 11, 12, 13] has mainly studied models from the HF Hub from the perspective of reusing deep pretrained models, the platform is not limited to hosting deep learning models only. Despite this, prior to conducting this research, we acknowledged that shallow ML models are less common. Based on the obtained data, the models were evaluated from an explainability perspective with regard to the employed class of ML algorithm and a three-tier categorization of model explainability, namely, non-explainable, mediocrely-explainable, and highly-explainable. Following prior work [3, 4, 5, 8, 9], models were analyzed in terms of two dimensions – performance and explainability. Our work introduces a third dimension, which considers the distribution and popularity of ML models in the real-world landscape, captured through models from the HF Hub. Hence, the main contributions of this work lie in the analysis of the distribution and popularity of the ML models shared on the HF Hub, contextualizing explainability research in the real-world usage landscape, thereby enhancing the practical relevance of XAI research.

The rest of the paper is organized as follows. **Section 2** provides an overview of the background and related work relevant to this work. **Section 3** details the research design utilized. The empirical results obtained based on the defined research design are presented and discussed in **Section 4**. Finally, **Section 5** synthesizes the findings and addresses threats to validity.

2. Background and Related Work

2.1. Explainability of Machine Learning Models

Explainability is related to the notion of explanation as an interface between a human – e.g., developer, theorist, ethicist, or end-user – and the decision-maker – the ML model –, where the explanation is an accurate approximation of the decision-maker and understandable to humans [5, 14]. The focus of this work is solely on end-users, i.e., domain experts, who use and rely on decisions, recommendations, or actions produced by a deployed intelligent system to assist their decision-making in work processes [3, 5]. As the focus of the explanation is on human understanding, empirical studies evaluate *perceived explainability*, which can serve as a determinant of the effectiveness of an intelligent system; when end-users can understand the behavior of the ML model used in an intelligent system, they are more likely to trust and act on its decisions, especially when those differ from their own expectations [5, 15].

In much existing XAI research, the explainability of individual ML models was conceptualized as a trade-off between performance and inherent explainability [3, 4, 8, 9]. For instance, simple models based on linear regressions were considered the most explainable, although they often do not perform as well. Conversely, high-performing models, such as those based on deep learning, were considered less explainable [3, 8, 9]. This initial view of the compromise between model performance and explainability is commonly graphically depicted in a two-dimensional Cartesian coordinate system, as presented in Figure 1 – subfigure A. Note that the coordinate axes are not quantified [16]. Theoretical contributions assumed an increase in the performance of ML models based on different ML algorithms combined with a continuous decrease in explainability; the relationship typically depicted as a linear or cubic curve [5]. The notion of ordering ML models based on the explainability of the underlying ML algorithm has been taken at face value for years, with little attempt at empirical confirmation. The first user-centered empirical research to gather evidence on this notion was conducted by Herm et al. [5]. The authors evaluated the end-user perceived explainability of ML models based on five common classes of ML algorithms, namely, linear regressions, decision trees, random forests, support vector machines, and deep neural networks, without utilizing any XAI augmentations. It is important to note that, although the research often refers to ensembles in general, we argue that it is more accurate to specifically refer to random forests. The investigation was conducted exclusively on random forests and due to the inherently interpretable nature of decision trees, their findings may not generalize to all

ensembles; thus, we opted for a more conservative approach to the naming of that class. The results of the research led to an updated visualization, as depicted in Figure 1 – subfigure B. Based on these findings, the authors grouped ML models based on explainability into three categories: *no explainability* (deep neural networks), *mediocre explainability* (linear regressions, support vector machines), and *high explainability* (decision trees, random forests), with the former encompassing deep, and the latter two shallow models [5].

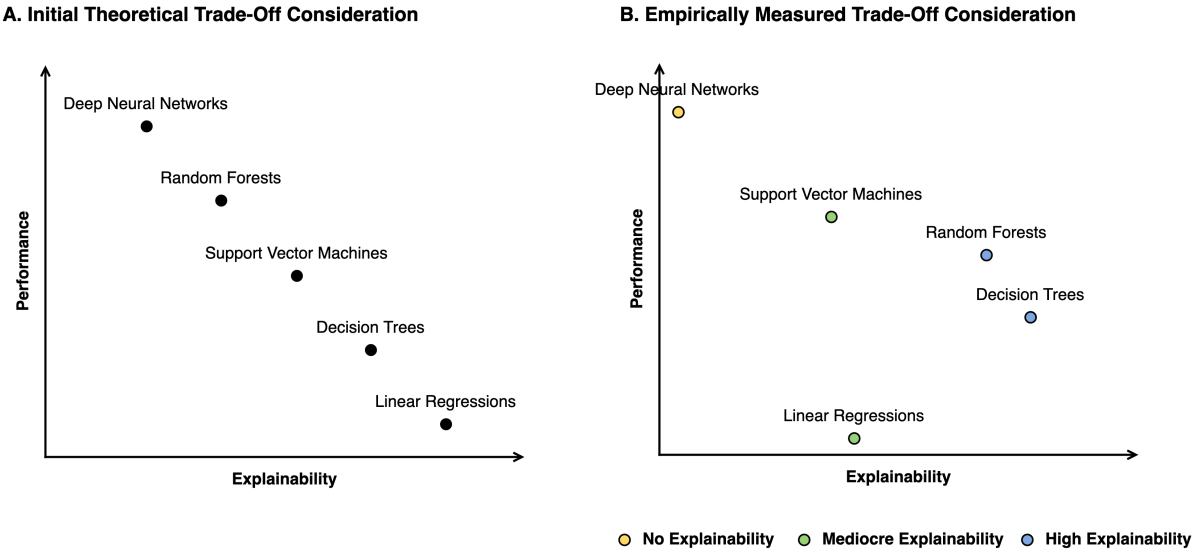

Figure 1: Previous **A. initial theoretical** and later **B. empirically measured** trade-off consideration between model performance and explainability based on the research by Herm et al. [5].

2.2. The Hugging Face Hub

General-purpose version-controlled hosting platforms, such as GitHub, offer means for collaboration on a common codebase and social interactions. The rise of ML led to the development of Git-based platforms specialized for ML-based projects [17]. The HF Hub [1] is a centralized Git-based repository hosting platform focused on ML-related content. It organizes its content into three registries: *models* (ML models shared by the community), *datasets* (datasets of different modalities – text, images, audio – and domains), and *spaces* (demonstrative interactive small-scale web applications) [1]. The platform is widely accepted by the ML community. This is demonstrated by the inclusion of the BERT, T5, and Gemma models shared by Google AI, the GPT-2 model shared by OpenAI, and the Llama model shared by Meta AI. Due to its popularity and utilization in practice, HF Hub has been previously analyzed as a source of state-of-the-art ML models, datasets, and projects. Previous works have focused on the reuse of pretrained models shared on the HF Hub [10, 13, 12], their evolution and maintenance [2], vulnerabilities [18], carbon footprints [19], and naming conventions [20]. Attempts to produce specialized solutions for analyzing the HF Hub community have also been made [17].

3. Research Design

3.1. Research Goal and Questions

Following the works of Castaño et al. [2, 19], a structured approach to defining research goals and questions based on the Goal-Question-Metric (GQM) methodology was employed. The primary research goal of this work was defined as *to analyze and evaluate ML models shared on the HF Hub to investigate the current state of perceived explainability from the perspective of end-users with regard to the distribution and popularity of these models across various common classes of ML algorithms.* To address this, two research questions (RQs) guided the research:

RQ1. *What is the distribution of different common classes of ML algorithms used in models shared on the HF Hub, and how does it relate to explainability?*

RQ2. *What is the distribution of different common classes of ML algorithms used in models shared on the HF Hub with regard to their popularity, and how does it relate to explainability?*

3.2. Data Mining Research

To address the research goal, a data mining research was conducted in three stages, namely data collection and preparation, categorization validation, and data analysis. An overview of the process is presented in Figure 2, and further discussed in the following subsections.

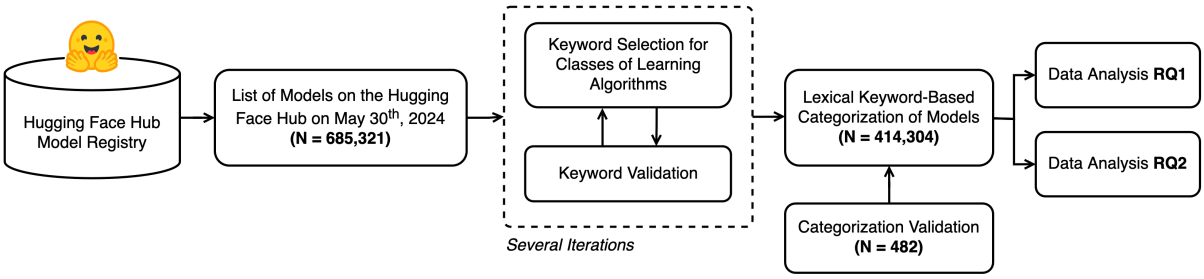

Figure 2: An overview of the data mining research process followed in this work.

3.2.1. Data Collection and Preparation

In the data collection and preparation stage, first, a comprehensive list of models available on the HF Hub was retreived using the HF API [21]. This retrieval was performed through the `HfApi` class and its `list_models()` function. For each model, the following data attributes were collected: `model_name` (the model's name), `tags` (the list of tags associated with the model), `created_at` (the creation date of the model's repository on the HF Hub), and `downloads` (the number of model's downloads in the last 30 days). Note that the earliest creation date recorded is March 2, 2022, 23:29:04 GMT, marking the start of HF Hub's creation date storage. Data collection was conducted on May 30th, 2024, resulting in a dataset of 685,321 models.

Since the specific ML algorithms used for each model are not directly reported, a lexical keyword-based model was employed to categorize each model based on its documentation (i.e., name and tags) into five common classes of ML algorithms as discussed by Herm et al. [5]. Initially, names and tags were converted to lowercase, and special characters (e.g., dashes, underscores) and whitespaces were removed. A regular expression approach was then used to match words in names or tags against predefined keywords related to each class. This approach is supported by previous findings by Jiang et al. [20], who demonstrated that model architectures could often be inferred from their names due to naming conventions on the HF Hub; in a sample of 300 randomly selected models from the HF Hub, 59% included information about their architecture in their names. To develop the predefined set of keywords, the first author conducted several iterations of keyword selection and validation, based on frequency analysis of *n*-grams of initial names and tags, as well as heuristics from existing body-of-knowledge. For instance, keywords for deep neural networks were also derived with the help from HF Hub documentation [22]. Models of unknown class of ML algorithms or based on ML algorithms not represented in five observed classes were categorized as *"Unknown"*. The set of keywords used for model categorization is presented in Table 1. Using the obtained list of models and the keyword-based categorization model, 414,304 models (60.45%) were categorized.

Table 1
Keywords for categorization of HF Hub models based on the class of ML algorithms employed.

Class of ML Algorithms	Keywords
Linear Regressions	linearregression, linreg, linearreg
Decision Trees	decisiontree, dtree, dectree
Support Vector Machines	supportvectormachine, supportvector, svmachine, svm
Random Forests	randomforest, randomforestregressor, randomforestclassifier, randomizedforest, rfregressor, rfclassifier
Deep Neural Networks	deepneuralnetwork, dnn, neuralnetwork, deeplearning, bert, llama, gpt, electra, longformer, nlp, transformer, deepnn, deepnet, transformers, albert, align, altclip, audiospectrogramtransformer, autoformer, bark, bart, beit, bertgeneration, bigbird, bigbirdpegasus, biogpt, bit, blenderbot, blenderbotsmall, blip, blip2, bloom, bridgetower, bros, camembert, canine, chineseclip, chineseclipvisionmodel, clap, clip, clipvisionmodel, clipseg, clvp, codellama, codegen, cohere, conditionaldetr, convbert, convnext, convnextv2, cpmant, ctrl, cvt, data2vecaudio, data2vectext, data2vecvision, dbrx, deberta, debertav2, decisiontransformer, deformabledetr, deit, depthanything, deta, detr, dinat, dinov2, distilbert, donutswin, dpr, dpt, efficientformer, efficientnet, encodec, encoderdecoder, ernie, erniem, esm, falcon, fastspeech2conformer, flauert, flava, fnet, focalnet, fsmt, funnel, fuyu, gemma, git, glpn, gptsw3, gpt2, gptbigcode, gptneo, gptneox, gptneoxjapanese, gptj, gptsanjapanese, graphormer, groundingdino, groupvit, hubert, ibert, idefics, idefics2, imagegpt, informer, instructblip, jamba, jetmoe, jukebox, kosmos2, layoutlm, layoutlmv2, layoutlmv3, led, levit, lilt, llava, llavanext, longformer, longt5, luke, lxmert, m2m100, mamba, marian, markuplm, mask2former, maskformer, maskformerswin, mbart, mctct, mega, megatronbert, mgpstr, mistral, mixtral, mobilebert, mobilenetv1, mobilenetv2, mobilevit, mobilevitv2, mpnet, mpt, mra, mt5, musicgen, musicgenmelody, mvp, nat, nezha, nllbmoe, nougat, nystromformer, olmo, oneformer, openllama, openaigpt, opt, owlv2, owlvit, paligemma, patchtsmixer, patchtst, pegasus, pegasusx, perceiver, persimmon, phi, phi3, pix2struct, plbart, poolformer, pop2piano, prophetnet, pvt, pvtv2, qdqbert, qwen2, qwen2moe, rag, realm, recurrentgemma, reformer, regnet, rembert, resnet, retribert, roberta, robertaprelayernorm, rocbert, roformer, rwkv, sam, seamlessm4t, seamlessm4tv2, segformer, seggpt, sew, sewd, siglip, siglipvisionmodel, speechencoderdecoder, speechtotext, speechtotext2, speecht5, splinter, squeezebert, stablelm, starcoder2, superpoint, swiftformer, swin, swin2sr, swinv2, switchtransformers, t5, tabletransformer, tapas, timeseriestransformer, timesformer, timmbackbone, trajectorytransformer, transfoxl, trocr, tvlt, tvp, udop, umt5, unispeech, unispeechsat, univnet, upernet, van, videollava, videomae, vilt, vipllava, visionencoderdecoder, visiontextdualencoder, visualbert, vit, vithybrid, vitmae, vitmsn, vitdet, vitmatte, vits, vivit, wav2vec2, wav2vec2bert, wav2vec2conformer, wavlm, whisper, xclip, xglm, xlm, xlmprophetnet, xlmroberta, xlmrobertaxl, xlnet, xmod, yolos, yoso

3.2.2. Categorization Validation

To validate the approach to model categorization and determine the error rate, a manual categorization of a sample of the models was performed. To determine the sample size, Cochran's sample size formula was considered, formally defined as:

$$n_0 = \frac{Z^2 \cdot p \cdot (1-p)}{E^2} \tag{1}$$

where n_0 is the sample size, Z is the selected Z-score corresponding to the desired confidence level (e.g., 1.96 for a 95% confidence level), p is the estimated proportion of an attribute present in the population, and E is the desired level of precision (i.e., margin of error) [23]. When dealing with a finite population, the formula can be adjusted as:

$$n = \frac{n_0}{1 + \frac{n_0 - 1}{N}} \tag{2}$$

where N is the size of the population, n_0 is the initial Cochran's sample size recommendation, as defined in Equation 1, and n is the adjusted sample size for a finite population [23]. The validation was conducted on a sample size of $n = 384$. Due to the acknowledged unbalanced nature of classes in HF Hub (favoring deep learning models), stratified sampling was utilized to extend the validation to four less-represented classes, increasing the sample size to $n = 482$.

Based on the obtained sample size, the first author manually categorized the models into five classes of ML algorithms and an additional *"Unknown"* class. For the categorization of each model in the

sample, the human annotator thoroughly reviewed the entire model documentation beyond the model name and tags, including the model card and, where available, the source code. The comparison between predicted classes by the automated categorization model and manually categorized classes is presented in Figure 3 in the form of a confusion matrix. It highlights the automated approach's adequate prediction ability. There are some miscategorizations between the *"Deep Neural Networks"* and *"Unknown"* classes, most can be attributed to additional data (i.e., model card and codebase) that the manual annotator had access to. Many miscategorizations related to the *"Support Vector Machines"* class were also attributed to models with long randomized name strings, which sometimes included *svm* abbreviations. To measure inter-annotator agreement between manual and automated model categorization, Cohen's Kappa (κ) was computed, which yield $\kappa = 0.867$, indicating an *Almost perfect* level of agreement (with the threshold for *Almost perfect* being $\kappa = 0.820$ per [24]).

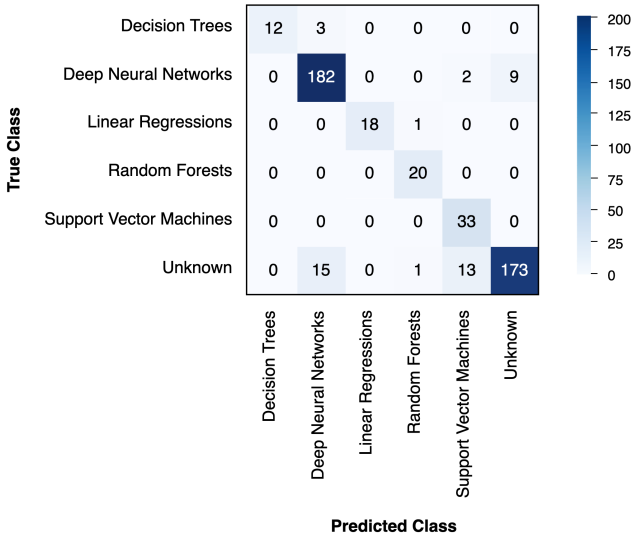

Figure 3: Confusion matrix comparing predicted classes by the automated categorization model against manually categorized classes, i.e., true class.

3.2.3. Metrics and Data Analysis

In the context of **RQ1**, the raw count of models for each class of ML algorithms was used. In the context of **RQ2**, the popularity of each class was estimated via the proxy of the number of downloads, as used in prior work [12]. This was done by multiplying the number of models associated with each class by their respective download counts.

4. Results and Discussion

The results of the categorization based on five common classes of learning algorithms are presented in Table 2. It can be observed that models utilizing deep neural networks represent the vast majority (~99.9%). Decision tree-based ($n = 13$) and linear regression-based models ($n = 18$) were published exceptionally rarely on the HF Hub platform. A graphical visualization depicting the trade-off between model performance and explainability in the context of model distribution on the HF Hub is presented in Figure 4 – subfigure A. To enhance readability, we represented the results using the size of the circles to indicate the model count instead of using a three-dimensional Cartesian coordinate system. The figure highlights the disproportionate frequency of models built on deep neural networks compared to all other observed algorithms. Furthermore, when accounting for the popularity of models, the disparity becomes even more apparent, as observed in Figure 4 – subfigure B.

The results of the categorization, considering a three-tier classification of models based on their explainability, are presented in Table 3. It can be observed that models with no inherent explainability,

Table 2
Distribution and popularity of models from the HF Hub categorized into five classes of ML algorithms.

	Model Distribution		Model Popularity	
Class of ML Algorithms	**Count**	**Percentage**	**Value**	**Percentage**
Linear Regressions	18	0.0043%	12	<0.0001%
Decision Trees	12	0.0029%	91	<0.0001%
Support Vector Machines	48	0.0116%	525	<0.0001%
Random Forests	22	0.0053%	32	<0.0001%
Deep Neural Networks	414,204	99.976%	1,261,748,276	99.999%

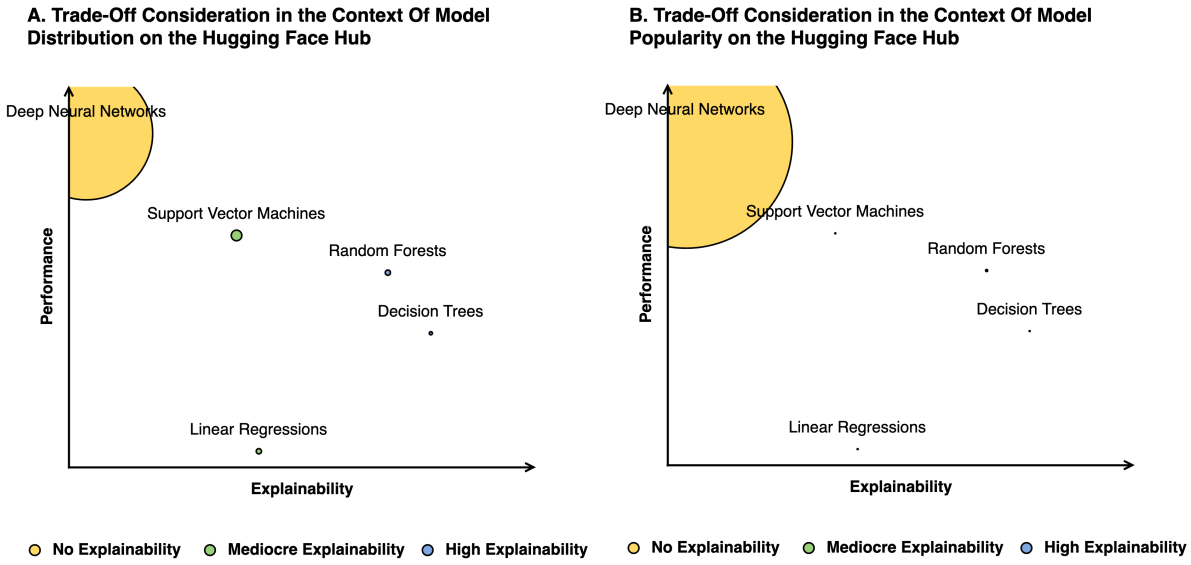

Figure 4: Trade-off consideration between model performance and explainability based on the research by Herm et al. [5] in the context of model **A. distribution** and **B. popularity** on the HF Hub.

as perceived by end-users, represent the vast majority (~99.9%), followed by models with mediocre explainability, and lastly, models with high explainability. An insight into the creation time of the models hosted on the HF Hub, presented in Figure 5, further illustrates the ongoing trend of prevalent deep neural network models with no inherent explainability over the months. Note that only models created after March 2022, the month when the HF Hub began storing model repository creation dates, are included in the figure. Additionally, for better readability, the values on the y-axis are displayed on a symmetric logarithmic scale.

Table 3
Model distribution and popularity of models from the HF Hub based on a three-tier explainability categorization.

	Model Distribution		Model Popularity	
Category of Model Explainability	**Count**	**Percentage**	**Value**	**Percentage**
No Explainability	414,204	99.976%	1,261,748,276	99.999%
Mediocre Explainability	66	0.0159%	537	<0.0001%
High Explainability	34	0.0082%	123	<0.0001%

In the context of **RQ1**, we observe that ML models based on deep neural networks, which offer no inherent explainability, prevail on the HF Hub over shallow ML models. This indicates that models might prioritize performance over explainability. This trend is even more apparent when considering their popularity (**RQ2**). However, as discussed in [4, 5, 16], it is important to emphasize that this is merely a general observation, as ordering ML models based on their underlying algorithm is hardly

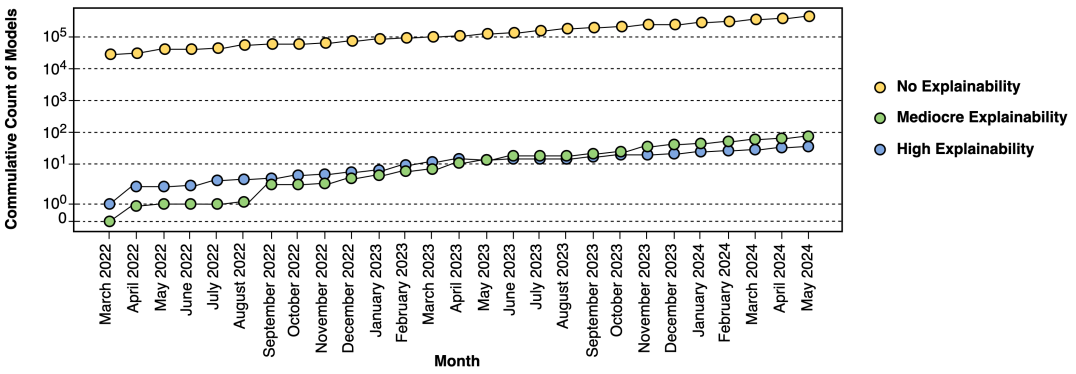

Figure 5: Cumulative count of models shared on the HF Hub per month based on a three-tier explainability categorization.

deterministic. More complex models do not necessarily achieve better performance, especially in domains with well-structured data, low data complexity, and limited data availability. Despite this, the dominance of deep neural network-based models and their predominantly black-box nature poses significant challenges in understanding and interpreting model decisions. This raises important user considerations, particularly in domains where transparency and accountability are crucial. Currently, challenges related to explaining deep neural network models are mostly associated with their model-free architecture and often lack adequate explanations of their full architecture (i.e., global explanations, as most approaches are local or visual, focused on a single instance at a time). Available global explanation methods are also computationally expensive due to the high number of features and parameters in the deep architectures of neural networks. Approximation models (created with methods such as LIME) for deep neural networks might closely resemble the black-box model for one dataset subset, while for other subsets, they might deviate significantly. Additionally, no global explanation method for text-based datasets can explain the rationale of deep neural network models [25]. Without adequate explanations of model behavior, users may hesitate to trust these models in critical applications. These challenges are compounded by the clear prevalence of deep neural network-based models on platforms like HF Hub. However, by acknowledging this trade-off and focusing on the development of explainability techniques tailored for deep neural networks, such as layer-wise relevance propagation and attention mechanisms, the research community can work towards reducing the disparity between performance and explainability. This will ultimately advance the responsible adoption of deep learning models in real-world scenarios. This agenda also aligns with DARPA's anticipation [3] of improvements in the explainability (and performance) of future techniques compared to the current state-of-the-art.

5. Conclusions

This paper addresses the explainability of ML models in real-world usage by analyzing data from the HF Hub platform. It evaluates shared models and their underlying ML algorithms concerning inherent end-user perceived explainability. Building on prior works [3, 4, 5, 8, 9], this work examines five classes of ML algorithms and a three-tier categorization of model explainability. The findings reveal a significant imbalance, with deep neural network-based models, which lack inherent explainability, dominating the platform. We highlight the need for future work, which should build on enhancing understanding of perceived explainability (as a trade-off to performance) beyond the five observed algorithm classes, ideally providing further insights into various deep neural network-based models and aligning algorithm categorization with ML field. Future work should also address further development of augmented XAI techniques specifically for deep neural network-based models, in line with their extensive usage and popularity, to ensure their transparent application across various domains.

5.1. Limitations and Threats to Validity

The data may contain inaccuracies, as ML algorithms employed are not explicitly defined for each model on the HF Hub, limiting our analysis to models with identifiable ML algorithm classes and introducing selection bias. The automated model categorization relied solely on model names and tags. Future studies might expand to other model documentation, e.g., model cards (though it should be noted that recent research reported that only half of the models have model cards [13]), and codebases. To assess the automated categorization approach, we performed manual validation on a sample of models, showing a small percentage of miscategorization and almost perfect agreement. We focused on five classes of ML algorithms for which end-user perceived explainability has been empirically analyzed by Herm et al. [5], potentially overlooking other relevant algorithms. It is essential to acknowledge the categorization used in the XAI field, and thus in [5], is not aligned with other common categorizations in the ML domain [26]. However, it was followed due to its common references in XAI literature. Additionally, some classes are narrow (e.g., linear regressions), while others are broader (e.g., deep neural networks). Relying solely on empirical evidence from this research increases the risk to the validity of our results, as data might not account for variability in explainability across different datasets, domains, and performance measures, as the research was conducted considering only two datasets (one with low data complexity, i.e., tabular dataset, and one with high data complexity, i.e., image dataset) from the healthcare sector, and considered performance in terms of accuracy. It should also be considered that the research focused solely on inherent explainability without considering additional post-hoc XAI augmentations. For measuring popularity, we used the number of downloads, a proxy used in prior work to indicate usage concentration [12]. However, future work might consider other measures, e.g., the number of likes and downstream reuse. Furthermore, the model registry from the HF Hub platform may not necessarily represent what is used to address real-world problems in general; it should be considered that the platform hosts predominantly pretrained models focused on natural language processing tasks [19]. Additionally, the results cannot be generalized to other ML model hubs, such as ONNX Model Zoo and PyTorch Hub. Though the results still serve as a trend indicator of ML models shared within the HF Hub community. The models were collected in May 2024; hence, the results might not align with future platform developments, especially due to rapid increases in shared models. This can be demonstrated by significant differences in available models in related works; a research carried out in February 2023 obtained around 110k models [18], in March 2023 170k models [19], and in November 2023 380k models [2], while we collected over 685k models. To mitigate this, we have detailed our approach to ensure it can be replicated in the future.

Acknowledgments

The authors acknowledge the financial support from the Slovenian Research Agency (Research Core Funding No. P2-0057).

References

[1] Hugging Face, Hugging face hub documentation, 2024. URL: https://huggingface.co/docs/hub/index, Accessed: 2024-06-03.

[2] J. Castaño, S. Martínez-Fernández, X. Franch, J. Bogner, Analyzing the Evolution and Maintenance of ML Models on Hugging Face, in: 2024 IEEE/ACM 21st International Conference on Mining Software Repositories, 2024, pp. 607–618.

[3] D. Gunning, D. Aha, Darpa's explainable artificial intelligence (XAI) program, AI magazine 40 (2019) 44–58.

[4] A. Barredo Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, et al., Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI, Information Fusion 58 (2020) 82–115.

[5] L.-V. Herm, K. Heinrich, J. Wanner, C. Janiesch, Stop ordering machine learning algorithms by their explainability! A user-centered investigation of performance and explainability, International Journal of Information Management 69 (2023) 102538.
[6] European parliament, EU AI Act: First Regulation on Artificial Intelligence, https://europarl.europa.eu/news/en/headlines/socwww.iety/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence, 2023. Accessed: 2024-06-13.
[7] Office of Science and Technology Policy, Blueprint for an AI Bill of Rights, 2022. URL: https://whitehouse.gov/ostp/ai-bill-of-rights/, Accessed: 2024-06-19.
[8] H. K. Dam, T. Tran, A. Ghose, Explainable software analytics, in: Proceedings of the 40th International Conference on Software Engineering, 2018, pp. 53–56.
[9] S. Nanayakkara, S. Fogarty, M. Tremeer, K. Ross, et al., Characterising risk of in-hospital mortality following cardiac arrest using machine learning: A retrospective international registry study, PLoS medicine 15 (2018).
[10] W. Jiang, N. Synovic, M. Hyatt, T. R. Schorlemmer, et al., An Empirical Study of Pre-Trained Model Reuse in the Hugging Face Deep Learning Model Registry, in: 2023 IEEE/ACM 45th International Conference on Software Engineering, 2023, pp. 2463–2475.
[11] W. Jiang, N. Synovic, P. Jajal, T. R. Schorlemmer, et al., Ptmtorrent: A dataset for mining open-source pre-trained model packages, in: 2023 IEEE/ACM 20th International Conference on Mining Software Repositories (MSR), 2023, pp. 57–61.
[12] J. Jones, W. Jiang, N. Synovic, G. K. Thiruvathukal, et al., What do we know about Hugging Face? A systematic literature review and quantitative validation of qualitative claims, 2024. arXiv:2406.08205.
[13] M. Taraghi, G. Dorcelus, A. Foundjem, F. Tambon, et al., Deep Learning Model Reuse in the HuggingFace Community: Challenges, Benefit and Trends, 2024. arXiv:2401.13177.
[14] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, et al., A survey of methods for explaining black box models, ACM Computing Surveys 51 (2018) 1–42.
[15] B. Berger, M. Adam, A. Rühr, A. Benlian, Watch me improve—algorithm aversion and demonstrating the ability to learn, Business & Information Systems Engineering 63 (2021) 55–68.
[16] C. Rudin, Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead, Nature Machine Intelligence 1 (2019) 206–215.
[17] A. Ait, J. L. C. Izquierdo, J. Cabot, HFCommunity: A Tool to Analyze the Hugging Face Hub Community, in: 2023 IEEE International Conference on Software Analysis, Evolution and Reengineering, 2023, pp. 728–732.
[18] A. Kathikar, A. Nair, B. Lazarine, A. Sachdeva, et al., Assessing the Vulnerabilities of the Open-Source Artificial Intelligence AI Landscape: A Large-Scale Analysis of the Hugging Face Platform, in: 2023 IEEE International Conference on Intelligence and Security Informatics, 2023, pp. 1–6.
[19] J. Castaño, S. Martínez-Fernández, X. Franch, J. Bogner, Exploring the Carbon Footprint of Hugging Face's ML Models: A Repository Mining Study, in: 2023 ACM/IEEE International Symposium on Empirical Software Engineering and Measurement, 2023, pp. 1–12.
[20] W. Jiang, C. Cheung, M. Kim, H. Kim, et al., Naming Practices of Pre-Trained Models in Hugging Face, 2024. arXiv:2310.01642.
[21] Hugging Face, Hugging Face Hub API, 2024. URL: https://huggingface.co/docs/huggingface_hub/v0.21.4/en/package_reference/hf_api, Accessed: 2024-06-13.
[22] Hugging Face, Transformers: Autoconfig, 2024. URL: https://huggingface.co/docs/transformers/v4.41.3/en/model_doc/auto#transformers.AutoConfig, Accessed: 2024-06-19.
[23] W. G. Cochran, Sampling techniques, John Wiley & Sons, 1977.
[24] M. L. McHugh, Interrater reliability: The kappa statistic, Biochemia medica 22 (2012) 276–82. PMID: 23092060.
[25] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, L. Liu, Explaining deep neural networks: A survey on the global interpretation methods, Neurocomputing 513 (2022) 165–180.
[26] C. C. Aggarwal, Data Classification: Algorithms and Applications, 1st ed., Chapman & Hall/CRC, 2014.