DOI 10.5162/iCCC2026/P30

Möbius-Shapley: Native Feature Attribution for Quantum Logic Decision Trees

Samar Dhahbi
Chair of Database and Information Systems
Brandenburg University of Technology
Cottbus, Germany
dhahbsam@b-tu.de

Ingo Schmitt
Chair of Database and Information Systems
Brandenburg University of Technology
Cottbus, Germany
schmitt@b-tu.de

***Abstract*—The demand for transparent AI has made explainability crucial, particularly in high-stakes domains like healthcare and finance. Quantum Logic Decision Trees (QLDTs) offer interpretable classification but remain challenging for non-experts of logic to understand. Methods like SHAP provide model-agnostic explanations. However, there is a mismatch between the minterm-based QLDT's logic of feature negation and SHAP's assumption of feature missingness. We propose Möbius-Shapley, a native local (i.e., instance-specific) explanation method that bridges this gap by applying the Shapley value framework directly to Möbius-transformed QLDT weights. Our approach eliminates the semantic mismatch and provides native explanations, though we note that the exact derivation is computationally intensive for high-dimensional datasets compared to approximation methods.**

***Index Terms*—Explainable AI, Quantum Logic Decision Trees, Shapley Values, Möbius Transform, Feature Attribution**

I. Introduction

SHAP [1] (SHapley Additive exPlanations) has emerged as one of the most popular methods in explainable AI. Its game-theoretic framework for feature attribution bears structural similarity to Quantum Logic Decision Trees [2] (QLDTs). Both analyze feature combinations. However, a critical semantic mismatch exists: standard SHAP treats QLDTs as black boxes, approximating feature contributions by marginalizing over "absent" features. This contradicts the underlying minterm structure of QLDTs, where features are never missing but negated by the means of quantum logic.

This mismatch motivates our key insight: rather than applying generic SHAP, we can directly compute Shapley values from QLDT's native minterm representation. We introduce Möbius-Shapley, which transforms minterm weights to Möbius weights in order to avoid negation. This transformation enables faithful Shapley-based local (i.e., instance-specific) explanations tailored to QLDT's quantum logic semantics. Our approach resolves SHAP's "cancellation effect" of conflicting interactions and prevents the "impossible data" problem arising from standard marginalization. Möbius-Shapley approach shift in methodology fundamentally alters the nature of the explanation: whereas standard SHAP asks: "What makes this prediction different from the average case?" (a comparative analysis against a background), Möbius-Shapley asks: "How is this specific decision derived?" (a constructive analysis of a logic expression).

II. Literature Review

A. QLDT Fundamentals

A QLDT [2] is a two-class classifier based on CQQL [3], a quantum logic approach. In contrast to fuzzy logic [4], in CQQL all laws of a Boolean algebra are respected. Its evaluation is based on sums and products on n normalized attribute values from the unit interval $[0, 1]$.

1) Logic Representation: The QLDT tree is a visual representation of a logic-based model, it is constructed from a CQQL [3] expression e in disjunctive normal form (DNF), which is a disjunction of minterms. Each minterm is a conjunction of all n atomic conditions c_j, where c_j is a unary condition on a value of an attribute j that appears either in its positive (c_j) or negated ($\neg c_j$) form. Since each condition has two possible states, there are 2^n possible minterms. For example, for two attributes where `age` stands for high age and `income` stands for high income, we have four minterms: ($\neg$`age` $\wedge$ $\neg$`income`), (`age` $\wedge$ $\neg$`income`), ($\neg$`age` $\wedge$ `income`), (`age` $\wedge$ `income`). Every logical expression e can be represented by use of minterm weights $\theta_i \in \{0, 1\}$ indicating whether a minterm i is active (`true` $\leftrightarrow$ 1) or not (`false` $\leftrightarrow$ 0):

$$e = \bigvee_{i=0}^{2^n-1} (minterm_i \wedge \theta_i)\,. \tag{1}$$

From a logic expression e, a QLDT ($qldt(e)$) is generated which represents a compact visualization of e in form of a kind of a decision tree. See Figure 1.

2) Evaluation: Unlike classical trees that follow one path down to a single leaf, the evaluation of a QLDT involves navigating from the root to all active leaves in parallel.

Let an object $o = \{a_1^o, a_2^o, \ldots, a_n^o\}$ where $a_j^o \in [0, 1]$ is the value of the j-th attribute a_j of o, and $[\cdot]^o \in [0, 1]$ denote the evaluation of a CQQL condition on o:

$$[minterm_i]^o = \prod_{j=1}^{n} [c_{ij}]^o. \tag{2}$$

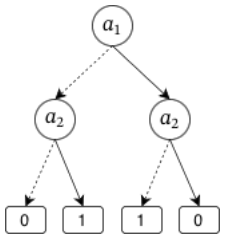
where c_{ij} refers to the j-th attribute condition c_j of $minterm_i$, if c_{ij} is positive then $[c_{ij}]^o := a_j^o$ otherwise $[c_{ij}]^o := 1 - a_j^o$.

Like the example shown in Figure 1, every minterm represents a path from the root to a leaf, where the right solid arrow indicates a positive condition and the left dashed arrow indicates its negation.

The results from all class-1-leaves are summed up to produce a class value in $[0, 1]$:

$$[e]^o = \sum_{i=0}^{2^n-1} \theta_i \prod_{j=1}^{n} [c_{ij}]^o =: [qldt(e)]^o. \quad (3)$$

A final threshold τ is applied to the result value from (3) for obtaining a discrete class decision (0 or 1).



$$e = (a_1 \wedge \neg a_2) \vee (\neg a_1 \wedge a_2)$$
$$[e]^o = [qldt(e)]^o = a_1^o \cdot (1 - a_2^o) + (1 - a_1^o) \cdot a_2^o$$

Fig. 1. A QLDT for an XOR example

B. SHAP

SHAP [1] (SHapley Additive exPlanations) adapts Shapley values [5] from cooperative game theory to machine learning explainability. In this framework, a model prediction is viewed as a game where the input features are the "players," and the specific prediction value is the "payout". For a set N of n features, the goal is to distribute the total "payout" (the difference between the model's prediction for a specific instance and the average prediction across the dataset) fairly among the features. The contribution of a feature j, denoted as φ_j, is calculated as the weighted average of its marginal contributions across all possible feature coalitions:

$$\varphi_j = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(n - |S| - 1)!}{n!} [v(S \cup \{j\}) - v(S)]. \quad (4)$$

where the core component of this calculation is the characteristic set function $v(S)$, which defines the model's output when a subset of features S is present. In the context of machine learning, features not in S (i.e., $N \setminus S$) are considered "missing". Since most models cannot handle missing inputs natively, SHAP approximates $v(S)$ via marginalization. It integrates over the missing features using a background dataset or an assumed distribution. SHAP holds the efficiency property, which guarantees that the sum of the feature attributions equals the total deviation of the prediction from the baseline:

$$v(N) - v(\varnothing) = \sum_j \varphi_j. \quad (5)$$

where $v(\varnothing)$ is the average prediction of background data.

III. Motivation

The structural similarity, combined with the critical semantic gap between minterms and the player set motivated our work to develop a native Shapley-based method for QLDTs without treating them as black boxes as it is done in the SHAP approach. By solely relying on the instance's native minterm representation rather than on average contribution, our method operates independently of background data distributions. Consequently, this leads to more transparent interpretation through a constructive breakdown of the specific decision.

A. The Structural Similarity

Interestingly, QLDTs inherently represent a form of coalitions, where each minterm evaluates a specific combination of feature presences and negations. This structure bears remarkable resemblance to the Shapley value framework, where the prediction-baseline difference (5) is decomposed into feature contributions.

B. Structure Incompatibility

Despite the similarity, there is a fundamental semantic gap. In minterms of quantum logic, all features are inherently present (positive or negated) with continuous truth values. The concept of "missing" or "absent" features in SHAP contradicts the foundational principles of minterms of a QLDT. In SHAP, when a feature is excluded from a coalition, it is treated as "missing" and marginalized over a background dataset. In minterms of a QLDT, features are never missing, negation $(1 - x)$ explicitly represents logical complement.

This gap means that while both frameworks compute contributions to feature combinations, they interpret "feature absence" differently. SHAP's marginalization approach may induce artificial scenarios. For example, in an exclusive-OR (XOR) rule, marginalizing over a missing feature can make it seem like the feature has no effect on average. This misleading cancellation can obscure the feature's critical role for a specific input. This weak point will be discussed in Section V.

IV. Methodology: The Möbius-Shapley Approach

We propose Möbius-Shapley, a native explanation method that bridges the gap between QLDT logic and Shapley analysis by utilizing the Möbius transformation.

A. Core Idea

Our approach leverages the unique structure of QLDTs to provide native explanations rather than treating the model as a black box. Our key insight recognizes that the Möbius transformation [6] can convert QLDT's minterm-based evaluation into a form compatible with Shapley analysis. By transforming minterms into feature interactions, we eliminate the negation/missingness mismatch while preserving all logical relationships.

B. Möbius Transformation

This mathematical transformation serves as the crucial translator between QLDT logic and Shapley analysis.

In our context, a minterm weight is interpreted as a value of set function $\theta_i = m(X)$ where $X \subseteq N$ is the set of features that appear positively in minterm i. Möbius transformation, like discussed in [7], maps $m(X)$ into interaction weights $\mu(X)$:

$$\mu(X) = \sum_{Y \subseteq X} (-1)^{|X|-|Y|} \cdot m(Y). \tag{6}$$

Let $\lambda_X(o)$ be the product of the feature values in the set X of an object o. Möbius transform maps the evaluation $[e]^o = \sum_{i=0}^{2^n-1} \theta_i \cdot [minterm_i]^o$ to an equivalent negation-free representation:

$$[e]^o = \sum_{X \subseteq N} \mu(X) \cdot \lambda_X(o). \tag{7}$$

For example, for an object $o = \{a_1, a_2\}$ instead of evaluating the expression $e = (a_1 \wedge \neg a_2)$ to $a_1 \cdot (1 - a_2)$, we compute it with a negation-free form as $[e]^o = \mu(\varnothing) \cdot 1 + \mu(\{a_1\}) \cdot a_1 + \mu(\{a_2\}) \cdot a_2 + \mu(\{a_1, a_2\}) \cdot a_1 \cdot a_2$.

For the expression $(a_1 \wedge \neg a_2)$, the Möbius weights computed from (6) yields:

$$\mu(\varnothing) = 0, \quad \mu(\{a_1\}) = 1, \quad \mu(\{a_2\}) = 0, \quad \mu(\{a_1, a_2\}) = -1$$

Substituting these values yields $[e]^o = 0 \cdot 1 + 1 \cdot a_1 + 0 \cdot a_2 + (-1) \cdot a_1 \cdot a_2 = a_1 - a_2 \cdot a_1$ which matches the original evaluation.

C. Method Overview

Our QLDT native feature attribution algorithm leverages the algebraic properties of the Möbius transform to solve the feature absence problem in QLDTs. The core innovation of this method lies in its formulation of the coalition value function $v(S)$ in (4). For a given instance o and a feature subset S, the value $v_o^{m\ddot{o}}(S)$ is defined as:

$$v_o^{m\ddot{o}}(S) = \sum_{X \subseteq S} \mu(X) \cdot \lambda_X(o). \tag{8}$$

This formulation replaces SHAP's marginalization approach, allowing us to compute contributions exactly from the model's interaction weights. Applying the Shapley formula in (4), we get the Möbius-Shapley values $\varphi_j^{m\ddot{o}}$ for a feature j:

$$\varphi_j^{m\ddot{o}} = \sum_{S \subseteq N \setminus \{j\}} \frac{|S|!(n - |S| - 1)!}{n!} [v_o^{m\ddot{o}}(S \cup \{j\}) - v_o^{m\ddot{o}}(S)]. \tag{9}$$

Our approach offers advantages over SHAP: it requires no background data, as the Möbius weights hold all necessary information. It also makes exact computations, eliminating the need for sampling approximation. Additionally, Möbius-Shapley method satisfies the efficiency axiom (5) where:

$$v_o^{m\ddot{o}}(N) - v_o^{m\ddot{o}}(\varnothing) = \sum_j \varphi_j^{m\ddot{o}}. \tag{10}$$

Based on (7) and (8), the full coalition value $v_o^{m\ddot{o}}(N)$ corresponds to the model evaluation $[e]^o$. For the empty set, equations (7) and (6) show that $v_o^{m\ddot{o}}(\varnothing) = \mu(\varnothing) = m(\varnothing) \in \{0, 1\}$. Since this term is instance-independent, we denote the constant baseline as $v^{m\ddot{o}}(\varnothing)$ thus $[e]^o = \sum_j \varphi_j^{m\ddot{o}} + v^{m\ddot{o}}(\varnothing)$.

The definition (10) marks a fundamental shift in baselines: While both methods utilize a constant reference point, standard SHAP derives its baseline **statistically** from the background dataset (the average prediction). In contrast, Möbius-Shapley derives its baseline **structurally** from the model itself (the weight of the all-negated minterm). Thus, SHAP measures deviation from the 'average case,' while Möbius-Shapley measures construction from the logical 'ground state'.

V. Results and Discussion

To validate the Möbius-Shapley framework, we applied the method to a QLDT modeling of an exclusive-OR (XOR) problem. This section details the feature attribution process and analyzes the advantages of this native approach compared to standard model-agnostic methods like SHAP.

A. A Real-World Example

We consider a cheese quality assessment based on storage conditions, modeled by two normalized features: `humidity` (H) and `temperature` (T). The model classifies a cheese as high quality (class 1) if the storage conditions are contrasting (e.g., high `humidity` with low `temperature`, or low `humidity` with high `temperature`). Conversely, it is classified as low quality (class 0) if both conditions are similarly high or low. This is modeled by a QLDT expression $e = (\neg \mathtt{H} \wedge \mathtt{T}) \vee (\mathtt{H} \wedge \neg \mathtt{T})$ (XOR) and a threshold $\tau = 0.55$. For a specific instance o where the cheese ring got stored in a cave with high `humidity` value ($\mathtt{H} = 0.7$) and low temperature ($\mathtt{T} = 0.3$), the QLDT evaluation is $[e]^o = (1 - 0.7) \cdot 0.3 + 0.7 \cdot (1 - 0.3) = 0.58$. Since its evaluation is higher than τ, this cheese ring is predicted as high quality (class 1). To explain this instance using Möbius-Shapley, first, we transform the QLDT minterm weights into interaction weights (μ) using Equation (6). This transformation reveals the feature interaction:

$$\begin{aligned}
\mu(\varnothing) &= m(\varnothing) = 0 \\
\mu(\{\mathtt{T}\}) &= m(\{\mathtt{T}\}) - m(\varnothing) = 1 - 0 = 1 \\
\mu(\{\mathtt{H}\}) &= m(\{\mathtt{H}\}) - m(\varnothing) = 1 - 0 = 1 \\
\mu(\{\mathtt{H}, \mathtt{T}\}) &= m(\{\mathtt{H}, \mathtt{T}\}) - m(\{\mathtt{H}\}) - m(\{\mathtt{T}\}) + m(\varnothing) \\
&= 0 - 1 - 1 + 0 = -2
\end{aligned}$$

Using the Möbius weights, we calculate the coalitional values according to Equation (8):

$$\begin{aligned}
v_o^{m\ddot{o}}(\varnothing) &= \mu(\varnothing) \cdot 1 = 0 \\
v_o^{m\ddot{o}}(\{\mathtt{H}\}) &= \mu(\varnothing) \cdot 1 + \mu(\{\mathtt{H}\}) \cdot \mathtt{H} = 0 + 1 \cdot 0.7 = 0.7 \\
v_o^{m\ddot{o}}(\{\mathtt{T}\}) &= \mu(\varnothing) \cdot 1 + \mu(\{\mathtt{T}\}) \cdot \mathtt{T} = 0 + 1 \cdot 0.3 = 0.3 \\
v_o^{m\ddot{o}}(\{\mathtt{H}, \mathtt{T}\}) &= \mu(\varnothing) + \mu(\{\mathtt{H}\}) \cdot \mathtt{H} + \mu(\{\mathtt{T}\}) \cdot \mathtt{T} \\
&\quad + \mu(\{\mathtt{H}, \mathtt{T}\}) \cdot \mathtt{H} \cdot \mathtt{T} \\
&= 0 + 0.7 + 0.3 + (-2) \cdot 0.21 = 0.58 = [e]^o
\end{aligned}$$

After applying Shapley formula (9), we obtain $\varphi_{\mathtt{H}}^{m\ddot{o}} = 0.49$ and $\varphi_{\mathtt{T}}^{m\ddot{o}} = 0.09$, which means that `humidity` contributes $+0.49$ of the total evaluation and `temperature` contributes $+0.09$.

B. Interpretation of Results

The results derived in the previous subsection align with the underlying XOR logic of the model in two ways:

1. Exclusivity: The calculated Möbius interaction weight of $\mu(\{\mathtt{H}, \mathtt{T}\}) = -2$ serves as a mathematical penalty. This negative value explicitly aligns with the XOR function, reducing the prediction value when both features are present simultaneously.

2. Dominance: For the specific input $(\mathtt{H} = 0.7, \mathtt{T} = 0.3)$ the condition $(\mathtt{H} \wedge \neg\mathtt{T})$ is the dominant active path. Feature `H` receives a higher attribution ($\varphi_{\mathtt{H}}^{m\ddot{o}} = 0.49$) because it strongly satisfies the positive condition, whereas Feature `T` contributes positively ($\varphi_{\mathtt{T}}^{m\ddot{o}} = 0.09$) but having a low value, effectively satisfying the negation $\neg\mathtt{T}$.

C. Comparative Analysis: Möbius-Shapley vs. Standard SHAP

While standard SHAP (e.g., KernelSHAP [1]) is effective for black box models, our analysis highlights that the native Möbius-Shapley approach offers distinct advantages by exploiting the QLDT structure. It eliminates marginalization, which prevents the "cancellation effect" and ensures the explanation is not derived from logically impossible data combinations.

1) Avoiding The Cancellation Effect: A critical limitation of standard SHAP is its tendency to obscure conflicting interactions by averaging them into a single attribution value. X. Huang and J. Marques-Silva [8] argue that marginalization can cause these positive and negative effects to cancel out, potentially resulting in near-zero Shapley values that imply feature irrelevance. For example, in the XOR scenario described in Subsection V-A, the features `humidity` and `temperature` are individually necessary but mutually exclusive. Regarding `humidity`: in a "high `temperature`" context, a storage cave being not humid might have a positive impact (+1), while in a "low `temperature`" context, it might be negative (-1). Standard SHAP averages these (0 contribution), obscuring the feature's importance. Möbius-Shapley avoids this by utilizing explicit interaction weights $\mu(\{\mathtt{H}, \mathtt{T}\}) = -2$, that act like a penalty when both features are present, providing a more "white-box" explanation.

2) Handling Feature Dependencies: The "Impossible Data" Problem: Standard SHAP methods typically assume feature independence when marginalizing over absent features. As Aas et al. [9] demonstrate, this assumption forces the model to average over impossible data combinations that don't make sense in real life. For example, when explaining a model that uses the features `sex` and `pregnant`, SHAP might create invalid data (e.g., `male` AND `pregnant`) to compute marginal contributions, which can lead to inaccurate explanations. In contrast, the Möbius-Shapley approach is native to the QLDT; it computes contributions solely from minterms weights, inherent in the model structure. By utilizing the explicit interaction weights (μ) derived from the Möbius transform, our method respects the Boolean dependencies without approximating missing features via unrealistic sampling.

3) Contrastive vs. Constructive Explanations: Both SHAP and Möbius-Shapley have inherent limitations. While SHAP's model-agnostic nature provides broad applicability, its reliance on marginalization can fail in QLDTs by causing semantic mismatches, cancellation effects, and impossible data scenarios. Möbius-Shapley attempts to resolve these issues by leveraging the native minterm representation of QLDTs, but it is computationally expensive due to the exponential number of minterms (2^n). While both methods employ Shapley values, they end up computing different values. Applying SHAP to the same example from Subsection V-A, we got: $\varphi_{\mathtt{H}} = 0.04$ and $\varphi_{\mathtt{T}} = 0.09$ with a base value $v(\varnothing) = 0.45$. Thus, they answer fundamentally different interpretative questions. Standard SHAP provides **contrastive explanations** that ask: "how does this prediction differ from the average?" by measuring feature contributions relative to the population baseline $v(\varnothing)$. In contrast, Möbius-Shapley provides **constructive explanations** that ask: "how is this specific prediction built?" by decomposing the actual evaluation starting from the all-negated baseline $v^{m\ddot{o}}(\varnothing)$. This distinction explains why the methods yield to different attribution values; they measure different quantities.

In our example: the cheese ring gets 0.58 quality score (prediction: $1 \leftrightarrow$ high quality).

TABLE I
SHAP VS. MÖBIUS-SHAPLEY EXPLANATION COMPARISON

Method	SHAP	Möbius-Shapley
Baseline	0.45 (global average)	0 (ground state)
`humidity`	+0.04	+0.49
`temperature`	+0.09	+0.09
Explanation	This cheese is high quality due to storage being warmer and slightly more humid than average	`humidity` is the main driver, with `temperature` providing secondary support

The question that arises is: If SHAP and Möbius-Shapley provide different values, then which of them are better for interpreting?

The answer is that both methods have complementary purposes. SHAP can be used to identify what makes a case unusual, while Möbius-Shapley can be used to understand how the model constructs the decision, revealing the internal reasoning and feature interactions.

VI. CONCLUSION

We have presented Möbius-Shapley, a native feature attribution method for Quantum Logic Decision Trees that resolves the fundamental semantic conflict between SHAP's "missingness" and QLDT's "negation". By leveraging the Möbius transformation, our approach derives non-approximated explanations directly from the model's structure, eliminating the

need for background data sampling and effectively addressing the “cancellation problem” in conflicting interactions. However, both approaches compute different contribution values, meaning that they answer different questions. SHAP is used for contrastive explanation whereas our approach provides constructive explanations.

Future work will focus on optimizing the computational efficiency of the Möbius transformation, particularly for high-dimensional datasets. We plan to investigate the application of this method to pruned QLDT to enhance scalability while maintaining interpretability.

REFERENCES

[1] S. M. Lundberg and S.-I. Lee: A Unified Approach to Interpreting Model Predictions, *Advances in Neural Information Processing Systems*, vol. 30, pp. 4765–4774, 2017.

[2] I. Schmitt, ”QLDT: A Decision Tree Based on Quantum Logic”, in *New Trends in Database and Information Systems – ADBIS 2022 Short Papers, Doctoral Consortium and Workshops*, 2022.

[3] I. Schmitt, ”QQL: A DB&IR Query Language”, *The VLDB Journal*, vol. 17, no. 1, pp. 39–56, 2008.

[4] L.mZadeh, ”Fuzzy sets”, *Journal of Information and Control*, 8, 1965.

[5] L. S. Shapley, ”A Value for n-Person Games”, *Contributions to the Theory of Games*, vol. 2, no. 28, pp. 307–317, 1953.

[6] G.-C. Rota, ”On the Foundations of Combinatorial Theory: I. Theory of Möbius Functions”, in *Classic Papers in Combinatorics*, pp. 332–360, Springer, 1964.

[7] G. Wirsching, I. Schmitt, and M. Wolff, ”Ausgewählte Anwendungen”, in *Quantenlogik Band 1*, Springer Vieweg, 2025.

[8] X. Huang and J. Marques-Silva, ”On the failings of Shapley values for explainability,” , in *International Journal of Approximate Reasoning*, , vol. 171, Art. no. 109112, Jan. 2024,

[9] K. Aas, M. Jullum, and A. Løland, ”Explaining Individual Predictions When Features Are Dependent: More Accurate Approximations to Shapley Values”, *Artificial Intelligence*, vol. 298, p. 103502, 2021.