Universiteit Leiden

Evaluating the Intrinsic Interpretability of Time-Conditioned Contraction Matching

Veronika Szekeres-Kovács (s4020065)

Supervisors:
Dr. Francesco Bariatti & Prof. Dr. Matthijs van Leeuwen

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl

July 15, 2026

Abstract

Time-Conditioned Contraction Matching (TCCM) is a deep anomaly detection method designed for tabular data. A key advantage of TCCM is its intrinsic interpretability: feature-wise importance scores originate from the model's internal structure [10]. As TCCM is a novel method, its interpretability (to the best of our knowledge) has not yet been studied in comparison to established post-hoc explanation methods such as SHAP. The goal of this thesis is to explore whether TCCM can provide explanations of comparable quality to established post-hoc methods and classical, interpretable machine learning algorithms while it maintains higher predictive performance.

This thesis investigates three research directions. Firstly, how similarly TCCM's intrinsic importance scores identify features contributing to anomalies compared with post-hoc explanations applied to the same TCCM model. Secondly, how do TCCM explanations compare with post-hoc explanations generated for other deep learning models? Thirdly, does TCCM achieve a better balance between accuracy and interpretability compared to classical interpretable machine learning approaches? To evaluate this, the methods are applied to large-scale tabular datasets. Then, feature-wise importance scores are normalized and ranked, and agreement between rankings are measured. The findings suggest that different explanation approaches may reveal different aspects of a model's decision-making. Therefore, intrinsic explanations should be considered together with established post-hoc explanation techniques, rather than its replacements.

Contents

1 Introduction

Context. The rapidly increasing volume and complexity of data in today's world makes it more challenging for humans to effectively collect, process, analyse, and interpret information. In the past decades, a need for these tasks to be outsourced has emerged to alleviate the pressure on employees and decision-makers. As a result, a variety of domains have adopted machine learning algorithms that identify patterns from large datasets with limited or no human intervention. Recent advances in deep learning, a subset of machine learning, have improved this ability even further, which led to a significant progress in several tasks, for example, in natural language processing, speech and image recognition, recommendation systems, and anomaly detection.

Anomaly Detection. The task of anomaly detection aims to identify instances or patterns in data that significantly deviate from other observations [1, 5]. Anomalies are often rare, yet they can provide valuable information about a system or a process. Consequently, anomaly detection has become an asset in many application domains, such as financial fraud detection, network intrusion detection, industrial fault monitoring, and medical diagnosis [1, 5].

As anomaly detection has started to adopt deep learning techniques, higher performance on large-scale or high-dimensional datasets has become more accessible compared to classical machine learning approaches. The core parts of deep anomaly detection methods that distinguish them from traditional machine learning are the highly complex neural networks that model relationships in big data. Anomalies can then be identified as observations that deviate from these learnt representations. While these methods enable strong predictive performance, they often have limited transparency due to the neural networks' underlying decision-making, which is difficult to interpret for humans [10]. Therefore, many deep anomaly detection methods are labelled as black-box models.

This lack of transparency poses challenges in fields where decisions must be understandable and justifiable. As anomaly detection may influence decisions that carry crucial consequences for people and organisations, these choices require clarity about why a certain instance was classified as an anomaly [10]. Consequently, there is a growing demand for anomaly detection methods that achieve a high performance while also providing meaningful explanations for their outputs.

Explainable Artificial Intelligence. This requirement ties into the field of explainable artificial intelligence (XAI), which presents methods that improve the interpretability of machine learning models while preserving their accuracy [2]. Conceptually, explainable artificial intelligence research differentiates between intrinsic interpretability and post-hoc explainability. Intrinsically interpretable models are interpretable by design [15], meaning that their decision-making process can be understood directly from the model structure. On the other hand, post-hoc explanation methods are applied on already trained models [15], usually to explain the predictions of black-box models. Two widely used post-hoc explanation methods are Local Interpretable Model-agnostic Explanations (LIME) [19] and SHapley Additive exPlanations (SHAP) [14]. Notably, SHAP has become widely used due to its theoretical foundations in game theory and its ability to provide both local and global explanations that are consistent with each other [14, 15].

Time-Conditioned Contraction Matching. Among recent developments in deep anomaly detection algorithms, Li et al. [10] introduced a new "scalable, explainable and provably robust"

anomaly detection method for tabular data called Time-Conditioned Contraction Matching (TCCM). The authors categorize their work as a deep, end-to-end, and semi-supervised approach where the model is trained only on normal instances and anomalies are identified during test time. The core mechanism of TCCM is that it learns the contraction path of normal instances in a residual velocity field, and it identifies anomalies as observations that deviate from this learnt contraction dynamics. More importantly, a compelling feature of TCCM is its intrinsic interpretability, which means that the importance scores that it produces originate from its learnt residual velocity field that marks deviations from the expected contraction behaviour. This attribute makes TCCM particularly interesting in the context of explainable anomaly detection.

Research Gap. The original TCCM study [10] showcases strong anomaly detection performance and attributes intrinsic interpretability. However, these explanations (to the best of our knowledge) have not yet been evaluated against established post-hoc explanation methods. More specifically, it is not yet explored to what extent TCCM's intrinsic feature-wise importance scores align with explanations generated by widely adopted techniques such as SHAP. Additionally, TCCM's position within the broader accuracy-interpretability trade-off has received limited attention in the literature. Many classical machine learning methods are considered easily interpretable but they may sacrifice predictive performance, while deep learning models usually achieve higher predictive performance at the expense of interpretability. Therefore, an important direction to explore is where TCCM lies on this spectrum.

Addressing these questions is meaningful for both researchers and decision-makers who rely on anomaly detection algorithms in practice. If TCCM, as an intrinsically interpretable deep anomaly detection model, can provide explanations of comparable quality to established post-hoc methods while it maintains higher predictive performance, it may reduce the need for computationally expensive explanations and improve the transparency of real-world anomaly detection systems.

Research Questions

To address the identified research gap, this thesis investigates the interpretability and predictive performance of TCCM through the following research questions:

- RQ1: To what extent do TCCM's intrinsic feature-wise importance scores identify the same features that contribute to an anomaly compared to post-hoc explanation methods applied on the same TCCM model?
- RQ2: How does TCCM's intrinsic interpretability compare to the explanations generated by post-hoc explanation methods applied on a different deep learning model?
- RQ3: Does TCCM achieve a better balance between accuracy and interpretability compared to classical interpretable machine learning approaches?

Approach and Contributions

Approach. To answer these questions, this thesis builds extensively on the original TCCM framework proposed by Li et al. [10]. The thesis compares intrinsic and post-hoc explanations by analysing the agreement between feature importance rankings produced by the different explanation

methods. Additionally, anomaly detection performance is also evaluated with established classification metrics. Lastly, the thesis contrasts TCCM with interpretable machine learning approaches, such as Decision Trees and Isolation Forest, in order to examine the trade-off between predictive performance and interpretability.

Contributions. The primary contribution of this thesis is a systematic evaluation of TCCM's intrinsic interpretability against established explainable models. Firstly, this work evaluates the agreement between TCCM's intrinsic feature-wise importance scores and SHAP explanations of the same TCCM model. Secondly, TCCM impportance scores are compared with SHAP explanations generated for an alternative deep anomaly detection model. Thirdly, the balance between predictive performance and interpretability is analysed by contrasting TCCM with classical interpretable machine learning approaches. With these contributions, the thesis provides new findings about the quality of TCCM's intrinsic explanations, their agreement with established post-hoc explanation methods, and the model's ability to balance predictive performance and interpretability.

Content. In the following sections, this manuscript first presents the theoretical background on anomaly detection, interpretability and explainability, and SHAP in Section 2. After this, Section 3 reviews related work about TCCM. Section 4 describes the research design, datasets and preprocessing, anomaly detection methods, explainability methods, and evaluation metrics used in this thesis. Next, Section 5 presents the results, and discusses the findings for each research question. Finally, Section 6 concludes the thesis, discusses its limitations, and outlines directions for future research.

2 Background

This section provides an overview of the most important theoretical concepts that this thesis builds upon. Firstly, it covers anomaly detection, particularly its challenges, anomaly detection settings, different anomaly scoring mechanisms, and deep anomaly detection. Secondly, it introduces the notions of interpretability, explainability, explainable artificial intelligence, and SHAP.

2.1 Anomaly Detection

The task of anomaly detection aims to identify instances or patterns in data that significantly deviate from other observations [1, 5]. These are called outliers, and they are mainly viewed as samples that do not fit the definition of normal behaviour, but in some cases they can be defined as instances that align with the conceptualized/laid out outlying behaviour [5]. In practice, outliers can occur in various forms, such as a cybersecurity attack, a credit card fraud, a mutation in genes, an automobile accident, or defects in machine components [5].

Challenges in Anomaly Detection. There are many challenges that are involved in anomaly detection tasks. Firstly, defining what the normal region is that encapsulates every normal instance can be difficult. For example, the boundary between normal and anomalous samples cannot always be clearly assigned, or the definition of normal behaviour can change throughout time. Secondly, the lack of labelled training and validation samples, and the similarity of anomalies to the noise in data both pose a challenge for anomaly detection. Thirdly, the definition of an outlier is problem-specific and varies across fields. [5]

Anomaly Detection Settings. Anomaly detection methods can be categorized based on how labelled instances are used: supervised, semi-supervised, or unsupervised. In supervised outlier detection, it is assumed that every instance in the training set is labelled, whether normal or anomalous. The advantage of this technique is that models have a clear representation of both classes. However, it may be challenging to or costly to acquire these labelled instances. The second type, semi-supervised anomaly detection uses labels for one class (usually the normal), precisely due to the difficulty of obtaining labels for the other class. In this approach, the anomalies become the samples that do not align with the modelled class. The last group, unsupervised anomaly detection models are commonly used as they do not require labelled data. On the other hand, the disadvantage of this technique is that it is prone to a higher false alarm rate. [5]

Anomaly Scoring. In terms of their outputs, outlier detection algorithms can be divided into labelling techniques and scoring techniques. While the former outputs two sets consisting of normal and anomalous instances, respectively, the latter provides an anomaly score for every observation. Consequently, the output of a scoring technique is a ranked list of instances with varying outlier scores. In this case, a threshold needs to be decided that marks the border between normal and anomalous samples, or the analyst can examine a selected number of top anomalies. [5]

Traditional Anomaly Detection. Those anomaly detection methods that do not rely on deep learning are called traditional anomaly detection approaches. These include methods such as Kernel Density Estimation, One-Class Support Vector Machine, or Isolation Forest [8].

Deep Anomaly Detection. Pang et al. [18] propose a categorization of deep anomaly detection methods with three paradigms: Deep Learning for Feature Extraction, Learning Feature Representations of Normality, and End-to-end Anomaly Score Learning. The models in Deep Learning for Feature Extraction derive low-dimensional representations of complex data in high-dimensional and non-linearly separable feature spaces, and the anomaly scoring is independent from the feature extraction. The second group of models, those that learn feature representations of normality, integrate feature learning and anomaly scoring within a unified framework. Lastly, end-to-end anomaly score learning differs from the previous approaches as it does not need to rely on predefined anomaly measures. Instead, neural networks are trained to directly produce anomaly scores with the help of a loss function.

2.2 Interpretability and Explainability

In order to gain an inisght into the working of deep anomaly detection models, it is important to use interpretable methods that are able to provide explanations about the original model's decision-making. In this thesis, the definitions by Roscher et al. [20] are used. Here, interpretability is the mapping of an abstract concept from the models into a domain that is understandable for humans, while explainability is "the collection of features of the interpretable domain, that have contributed for a given example to produce a decision" [20].

Explainable Artificial Intelligence. The field of explainable artificial intelligence (XAI) offers methods that improve the interpretability of machine learning models while preserving their accuracy [2]. Conceptually, explainable artificial intelligence research differentiates between intrinsic interpretability and post-hoc explainability. Intrinsically interpretable models are interpretable by design [15], meaning that their decision-making process can be understood directly from the model structure. On the other hand, post-hoc explanation methods are applied on already trained models [15], usually to explain the predictions of black-box models. Two widely used post-hoc explanation methods are Local Interpretable Model-agnostic Explanations (LIME) [19] and SHapley Additive exPlanations (SHAP) [14].

LIME and SHAP. Local Interpretable Model-agnostic Explanations (LIME) is a technique that approximates an interpretable model around the original model's prediction locally, resulting in precise and understandable explanations about the classifier's predictions [19]. The other method, SHAP has become widely used due to its theoretical foundations in game theory [14, 15]. This method marks the importance of every feature with a unified measure of feature importance (SHAP values) for a specific prediction [14]. An advantage of SHAP is that it can provide explanations for every instance (global), or for a selected instance (local), while LIME exclusively allows for local explanations [21]. On the other hand, SHAP is significantly slower than LIME [21]. Lastly, LIME cannot encapsulate nonlinear associations, but SHAP may be able to identify them [21].

3 Time-Conditioned Contraction Matching

This section summarizes the most important aspects and findings of the paper that introduced Time-Conditioned Contraction Matching (TCCM) [10] for the purpose of this thesis. TCCM is a new, "scalable, explainable and provably robust" anomaly detection method for tabular data [10]. The authors categorise their work as a deep, end-to-end, and semi-supervised approach.

Intuition. The core idea of TCCM is to model the behaviour of normal observations as a time-conditioned contraction vector with a fixed target. This vector shows how instances move towards the learnt regions of normal data at each sampled time step. In the originally proposed anomaly detection setting, the model is trained exclusively on normal samples to learn the residual velocity field that characterizes the expected contraction dynamics of normal instances. Those samples that align with the learnt contraction dynamics are considered normal, while the ones that are inconsistent with it are anomalous. The larger the residual is, the more likely it is that the model detected an anomaly.

Training. During training, the model aims to minimize a loss shown in Equation 1 (taken from [10]). Through this objective, the model learns the direction and magnitude of the time-dependent contraction dynamics. The velocity vector of normal points moves towards the origin and therefore approximates $-z$. Anomalous samples have larger residuals than normal points, which signals that the model never learnt this point's structure as they deviate from the expected contraction path.

$$\min_{\theta} E_{z\sim p_{\text{data}},\, t\sim\mathcal{U}(0,1)} \left[\left\| f_\theta\left([z; \text{Embed}(t)]\right) + z \right\|_2\right] \tag{1}$$

In this training objective, θ denotes the model parameters, z is sampled from the data distribution p_{data}, and t is sampled from the interval $(0, 1)$. $\text{Embed}(t_{\text{fixed}})$ is the sinusoidal embedding of the selected time step. This is concatenated with the input z and then passed into a multilayer perceptron. The notation $f_\theta([z; \text{Embed}(t_{\text{fixed}})])$ is the learnt velocity field that is conditioned on z and t. The function $f_\theta(\cdot)$ represents this neural network model which predicts a contraction vector. Finally, the $\mathcal{L}_2$ norm is denoted as $\|\cdot\|_2$.

Anomaly Scoring. Formally, the anomaly score of a test sample z evaluated at a fixed time step $t_{\text{fixed}} \in (0, 1]$ can be defined as in Equation 2 (taken from [10]).

$$S_{\text{fixed}}(z; t_{\text{fixed}}) = \left\| f_\theta\left([z; \text{Embed}(t_{\text{fixed}})]\right) + z \right\|_2 \tag{2}$$

The anomaly score of TCCM is derived from the residual components of the learnt velocity field. Generally, instances with misaligned velocities and larger residuals receive higher anomaly scores. Consequently, anomaly detection highlights the degree to which a sample deviates from the contraction dynamics learnt from normal data.

Interpretability. This model aims to overcome the lack of interpretability that is common among deep anomaly detection approaches. Notably, a key contribution of TCCM is its intrinsic interpretability. More specifically, the residual vector $f_\theta\left([z; \text{Embed}(t_{\text{fixed}})]\right) + z$ of an instance lives in the feature space, and the magnitude of each component directly indicates the contribution of

the corresponding feature to the anomaly score [10]. Consequently, TCCM can produce intrinsic feature-level importance scores that represents the underlying behaviour of the model. As these explanations originate directly from the model's internal computations, no additional explanation algorithms (e.g., post-hoc methods) are required.

Evaluation. The authors evaluate TCCM on multiple benchmark anomaly detection datasets from the ADBench suite [7] and compare its performance with several classical and deep learning-based baselines. Their results demonstrate that TCCM achieves state-of-the-art anomaly detection performance while it also provides intrinsic interpretability.

4 Methodology

This chapter provides an overview of the research design, justification of the selected methods, and the process of how the research questions are examined.

4.1 Research Design

This thesis aims to evaluate the explainability of Time-Conditioned Contraction Matching (TCCM) [10]. The main objective is to compare TCCM's intrinsic feature-level explanations against an established post-hoc explainability method and classical interpretable machine learning baselines. The three research questions are investigated in an experimental design, comparing two or more outcomes.

Experiment Design. Firstly, we compare TCCM's intrinsic explanations to SHAP importance scores obtained from the same TCCM model. This experiment provides insight into how TCCM's own explanations align with the feature importance scores produced by a post-hoc explanation method.

Secondly, we compare TCCM to another deep anomaly detection method called non-parametric Diffusion Time Estimation (DTE) [13]. The latter is selected due to its comparably high performance in the results of the original TCCM study [10]. The two approaches are applied on the same selected datasets; additionally, SHAP is applied on DTE. This setup enables the analysis of the differences between TCCM's intrinsic interpretations and post-hoc explanations in deep anomaly detection models.

Thirdly, we investigate the trade-off between accuracy and interpretability. Two classical interpretable machine learning models, Decision Tree and Isolation Forest, are compared against TCCM's predictive performance and feature ranking. These methods are chosen due to their interpretable nature and simplicity. The results of these experiments contribute to shaping TCCM's position in the accuracy-interpretability balance.

Data and Evaluation. All experiments use three of the ADBench tabular datasets [7] obtained from the original TCCM repository [9], and an additional publicly available dataset published by Dr. Nour Moustafa and Dr. Jill Slay from the Intelligent Security Group of UNSW Canberra [16].

Performance evaluation and explainability evaluation follow the same experimental procedure. Firstly, the models are trained, then the anomaly scores are generated (see Section 2 for a general description and Section 3 for a description of the TCCM anomaly scoring). Subsequently, the feature-wise importance scores are obtained, normalized, and averaged over the selected instances. Then, the average feature importance scores are ranked from highest absolute score to lowest absolute score to form feature rankings for each model. Lastly, these feature rankings are compared and evaluated.

Evaluation criteria are explained in more detail in Section 4.5, but in short, the anomaly detection performance is evaluated with AUROC. The evaluation of explainability is measured by ranking-based similarity measures, such as Jaccard similarity and Rank-Biased Overlap (RBO). Additionally, statistical significance tests are applied to assess whether observed explanation similarities differ meaningfully from random agreement.

The experiments are implemented in Python, and were carried out on the machines of the Leiden Institute of Advanced Computer Science Research and Education Lab (REL) cluster that can host CPU-intensive computations.

The next sections describe the datasets, preprocessing, anomaly detection and explainability approaches, and evaluation metrics used throughout the experiments.

4.2 Datasets & Data Preprocessing

In order to build on previous experiments conducted on TCCM [10], three tabular datasets from the open-source ADBench benchmark collection [7] are selected for the explainability and anomaly detection experiments. The experiments use the preprocessed ADBench versions of *celeba*, *cover*, and *magic.gamma* that are distributed with the official TCCM repository [9]. For these datasets, we reconstructed the feature names from the original dataset sources mentioned in this section.

These datasets aim to support both anomaly detection and explainability evaluation. As TCCM is a deep learning-based anomaly detection method, we chose datasets with a relatively large number of instances: the majority of the selected datasets contain more than 100,000 samples. In addition, to allow for a meaningful explainability comparison between intrinsic and post-hoc explanation methods, a sufficient number of features are required. Therefore, datasets with fewer than 10 features are excluded. A summary of the datasets can be found in Table 1.

Dataset Description. The *celeba* dataset originates from the CelebFaces Attributes (CelebA) dataset [12], where facial images are annotated with binary attributes. Each sample represents a picture and has 39 binary features such as hair color, face shape, makeup, and accessories. While the original CelebA dataset [12] does not have labelled anomalies, the ADBench version that the TCCM repository uses [9] is adjusted for an anomaly detection task. In the ADBench version of *celeba*, the "Bald" attribute has been removed and used as the anomaly label, where baldness is treated as an anomaly and non-bald individuals are treated as normal samples. This dataset is selected due to its large sample size and relatively high feature dimensionality.

The *cover* dataset is derived from the Covertype dataset from the UCI Machine Learning Repository [3]. The original dataset is designed for forest cover type classification based on cartographic data collected from wilderness areas in Northern Colorado. The ADBench/TCCM repository [9] version contains 10 numerical features, such as elevation, slope, hillshade, and distances to hydrology, roadways, and fire points. The anomaly detection task is formulated as identifying a certain forest cover type. This dataset is chosen due to its large sample size and easily interpretable context.

The *magic.gamma* dataset comes from the MAGIC Gamma Telescope dataset from the UCI Machine Learning Repository [4]. The dataset contains Monte Carlo-generated measurements that simulate the detection of high energy gamma particles using the imaging technique in a ground-based atmospheric Cherenkov gamma telescope. Instances are represented with 10 continuous physical features, such as length, width, concentration, and asymmetry. In terms of classes, gamma signal events and hadronic background events are separated, and hadronic events are treated as anomalies. This results in a significantly higher anomaly ratio (35.16%) than in the other selected datasets.

The fourth dataset used in this thesis is the UNSW-NB15 network intrusion detection dataset by Moustafa and Slay [16]. The dataset was generated using the IXIA PerfectStorm tool in the Cyber Range Labe of the Australian Centre for Cyber Security to simulate modern network traffic

that records both regular traffic and attacks [16]. Every record contains 49 attributes related to network connections, such as protocol information, packet statistics, timing, and TCP state information. The dataset contains two label columns representing the attack category, and the binary anomaly label. The original UNSW-NB15 dataset is available on one of the author's website [17], and is distributed across four CSV files. For this thesis, we selected the first file, which contains 700,001 samples, to introduce another large-scale dataset into the experiments. The binary "label" attribute is used as the anomaly detection target, where normal instances are labelled as 0 and attack samples as 1.

Data Preprocessing. The *UNSW-NB15* dataset require preprocessing to transform the mixed numerical and categorical network traffic features. Categorical attributes, including source IP address, destination IP address, transaction protocol, connection state, and service type, are converted to string format and encoded using one-hot encoding. In the case of IP addresses, there are only 40 unique source IP values and 44 unique destination IP values, therefore these values can be one-hot encoded as well instead of being dropped. Numerical attributes are converted to floating point values, while hexadecimal representations are transformed into decimal integers. We replace missing numerical values with zeros. After preprocessing, the numerical and one-hot encoded categorical features are combined into a sparse matrix that contains 290 features. The "attack_cat" column is excluded from the feature matrix, while the binary "label" attribute is retained as the ground-truth anomaly indicator for evaluation purposes.

Data	Samples	Features	Anomaly	% Anomaly	Category
celeba	202 599	39	4 547	2.24	Image
cover	286 048	10	2 747	0.96	Botany
magic.gamma	19 020	10	6 688	35.16	Physics
UNSW-NB15	700 001	290	22 215	3.17	Cybersecurity

Table 1: Summary of the preprocessed datasets used in the experiments.

According to the experimental setup in the original research [10], each dataset is divided into training and test sets following a semi-supervised anomaly detection setting. The normal samples are randomly divided into equally sized training and test subsets with a reproducible train-test split. In the first two research questions' experiments, half of the normal samples are assigned to the training set, while the remaining normal samples and all anomalous instances are included in the test set. Consequently, models are trained exclusively on normal observations and evaluated on mixed test data. The only deviation from this setup occurs in RQ3, where TCCM is compared against a Decision Tree.

In the RQ3 experiments, the datasets are divided using a stratified train-test split with 70% of the samples assigned to the training set and 30% to the test set. For TCCM and Isolation Forest, anomalous samples are removed from the training set after splitting, resulting in a setting where both methods are trained exclusively on normal instances. In contrast, the Decision Tree is trained on both normal and anomalous training samples. The test sets remain unchanged and contain both normal and anomalous instances. Finally, all the features in every experiment are standardized prior to training.

4.3 Anomaly Detection and Baseline Models

This subsection describes how the anomaly detection methods were constructed and setup for the experiments. For all experiments, a fixed random seed ensures reproducibility for dataset splitting, anomaly selection, and SHAP background sampling.

4.3.1 Time-Conditioned Contraction Matching

For the experiments conducted in this thesis, TCCM models are trained using the official repository implementation and the provided dataset-specific hyperparameter configurations [9]. For the *UNSW-NB15* dataset, the hyperparameters are specified manually in 10 training epochs, a batch size of 512, and a learning rate of 0.005.

First, the TCCM models are trained separately for every dataset. Then, anomaly scores for the test instances are computed with the model's original anomaly scoring function, followed by min-max normalization to the range of $[0, 1]$. TCCM's anomaly detection performance is evaluated with the AUROC computed from the raw anomaly scores and the ground truth test labels.

In addition to overall anomaly scores, TCCM can provide intrinsic feature-wise explanations. In order to obtain TCCM's intrinsic feature importance scores, these are reconstructed from the residual velocity field that the model's internal neural network had generated. The test samples are passed through the internal neural network at a fixed timestep $t = 1$, which produced the predicted contraction vectors. Then, these vectors are combined with the original input representation to form signed residual vectors. The absolute values of these residuals are interpreted as the feature-wise anomaly contributions. Lastly, to ensure comparability across instances, the feature importance values are normalized row-wise so that the contributions of all features for a single sample sum to one.

4.3.2 DTE Non-Parametric

Diffusion Time Estimation (DTE) is a deep anomaly detection method by Livernoche et al. [13]. DTE models the distribution of diffusion times for a given input, and derives the anomaly score from either the mean or the mode of the resulting distribution.

In this thesis, the non-parametric approach of DTE is used as it achieved a strong anomaly detection performance in the original TCCM study [10]. The second research question's experiment is conducted using the DTE non-parametric algorithm and the semi-supervised setting provided in the TCCM-NIPS repository [9].

As a start, the DTE non-parametric model is trained separately for every dataset using the normal training instances. After training, anomaly scores for the test instances are obtained with the model's original anomaly scoring function and evaluated with AUROC using the ground truth test labels. The resulting anomaly scores are min-max normalized to the range $[0, 1]$ to allow for comparisons with other methods. Higher anomaly scores indicate a stronger deviation from the learnt normal diffusion behaviour.

In the next phase, the anomaly ranking of DTE non-parametric is used to determine which instances would be explained by SHAP. To allow for a comparison with TCCM's intrinsic interpretability, post-hoc explainability is applied in the form of a SHAP KernelExplainer, which uses the DTE anomaly scoring function as the prediction target and estimates local feature contributions for selected anomalous test instances.

4.3.3 Decision Tree

Decision Tree is selected as a classical interpretable baseline for anomaly detection due to its transparent structure and inherent feature interpretability. The hierarchical tree structure allows for direct insight into which features contribute to classification outcomes.

As Decision Tree classifiers require labelled training data, the experimental setup for this comparison differs from the semi-supervised setup used for TCCM and DTE (see Section 4.2). The classifier is trained using both normal and anomalous instances, and with a fixed random seed to ensure reproducibility. To improve class balance, the classifier is defined with balanced class weight. The maximum tree depth is fixed to 4 and the minimum number of samples required in each leaf node was fixed to 50. Prior to training, all features are standardized using the same preprocessing that are applied in the other experiments.

After training, anomaly scores for the test instances are obtained from the predicted probability of the anomaly class, and the anomaly detection performance is evaluated with AUROC.

Feature importance is estimated with permutation feature importance. For each feature, the values in the test set are randomly shuffled while all other features are kept unchanged. The resulting decrease in AUROC is used as the feature importance score, and the procedure is repeated 10 times per feature. The average decrease across repetitions is used as the final importance estimate, while negative importance values are replaced with zero.

4.3.4 Isolation Forest

Isolation Forest is selected as a classical anomaly detection baseline as, opposed to Decision Trees, it does not require anomalies for training. Isolation Forest identifies anomalies through recursive random partitioning of the feature space [11].

For the Isolation Forest experiment, the model is trained separately for each dataset using the same normal training instances as TCCM. In terms of the Isolation Forest settings, the number of trees is fixed to 100 and the contamination parameter is set to `auto`.

After training, anomaly scores for the test instances are computed using the model's decision function that provides continuous anomaly scores. Since this function assigns higher values to more normal samples, the scores are multiplied by -1 so that larger values correspond to stronger anomalies, as in the case of TCCM. The resulting anomaly scores are evaluated with the performance metrics mentioned in Section 4.5.1, and min-max normalized to the range $[0, 1]$ to allow for comparisons with other methods.

Isolation Forest does not intrinsically provide feature-level explanations for individual predictions. Therefore, feature importance estimation is performed using permutation importance. The importance of a feature is determined by measuring the decrease in anomaly detection performance (AUROC) after randomly shuffling the feature values 10 times. The average decrease across repetitions is used as the final feature importance score. Negative importance values are changed to zero as they do not correspond to meaningful positive contributions to anomaly detection. Then, the resulting feature importance vectors are normalized in a way that the contributions summed to one. This allows for direct comparison with the normalized feature importance scores of TCCM. Those features whose permutation resulted in larger decrease in performance are interpreted as more important for the model's predictions.

4.4 SHAP

The explainable method used for the experiments of the first two research questions is SHapley Additive exPlanations (SHAP), a well-established post-hoc explanation model described in Section 2. Larger absolute SHAP values indicate features with greater influence on the anomaly score assigned by the model.

KernelExplainer. In this study, SHAP KernelExplainer is selected because it can directly explain anomaly scoring functions through a model-agnostic framework. Although TCCM and DTE are both neural network models, the objective of the experiments is to explain the final anomaly scores produced by the models rather than the internal neural network outputs. Consequently, the KernelExplainer is selected as it only requires a prediction function and therefore can provide a consistent explanation framework across all evaluated anomaly detection methods.

Experimental Setup. For prediction targets, SHAP KernelExplainer is applied to the trained models' (TCCM in RQ1 and DTE in RQ2) anomaly scoring functions. To reduce the computational cost of Kernel SHAP, a random subset of the normal training instances is used as the SHAP background distribution.

The instances selected for explanation are determined based on the anomaly ranking produced by TCCM in RQ1 and DTE in RQ2, respectively. First, test samples are sorted based on their anomaly scores in descending order, where higher scores indicated stronger anomalies. Secondly, a threshold needs to be defined in terms of what is considered as an anomaly in the ranking. Instead of defining an exact cutoff value, we take into account the number of true anomalies that are present in the test set, and select the same amount of samples from the sorted anomaly score list to be explained.

Due to the high computational complexity of the explanation process of the DTE anomaly scoring function, SHAP values are computed only for a selected subset of anomalous instances in the RQ2 experiment. These anomalies are obtained using stratified sampling from the detected anomaly set of DTE. The anomalies are divided into three equally sized groups according to their anomaly scores, representing low-, medium-, and high-scoring anomalies. The final subset is sampled proportionally from these groups in order to preserve the distribution of anomaly scores while reducing the number of instances requiring explanation. On the other hand, this restriction of high computational cost does not pertain to when SHAP is applied to the TCCM anomaly scoring function, therefore, all the anomalous instances are explained in the RQ1 experiment.

For the SHAP values, the number of perturbation evaluations per explained instance is fixed to 500 in RQ1, and 50 in RQ2. These values are set after experimenting with different values to find a reasonable runtime. As the explainer that is applied to DTE is significantly slower, the number of perturbation evaluations are fewer in RQ2. After obtaining the scores, absolute SHAP values are interpreted as feature importance scores and normalized row-wise such that the contributions of all features for a single sample summed to one. These normalized feature importances are subsequently used in the explainability agreement evaluation.

4.5 Evaluation Metrics

This subsection provides an overview of the evaluation metrics that are used to compare the performance and the explainability of the chosen methods.

4.5.1 Performance Metrics

To evaluate anomaly detection performance, we calculate the Area Under the Receiver Operating Characteristic (AUROC) directly from the anomaly scores. A Receiver Operating Characteristic (ROC) curve shows the trade-off between the true positive rate and the false positive rate, and it can be advantageous for datasets that have a skewed class distributions [6]. AUROC measures the area under the ROC curve and provides information about how well a model differentiates between classes. The AUROC score is between 0 and 1, where 1 marks a perfect performance, 0.5 can be interpreted as guessing, and smaller scores indicate a performance that is subpar to random guessing [10].

4.5.2 Explainability Metrics

To obtain feature rankings, the normalized feature importance scores are averaged across all explained anomalous instances. To evaluate the agreement between feature rankings, we use Jaccard similarity and Rank-Biased Overlap (RBO), while Spearman's rank correlation coefficient is used to test whether a statistically significant association exists between the rankings.

Jaccard Similarity. Jaccard similarity is calculated between two ranked feature lists at top k, where only the first k items of each ranking are considered and the order is ignored. The experiments evaluate Jaccard similarity at four ranking depths: top 1, top 3, top 5, and top 10. The metric is computed as the ratio of the number of shared features to the total number of unique features appearing in the two top k lists. The resulting values range from 0 to 1, where 0 indicates no overlap between the rankings and 1 indicates identical feature sets. As Jaccard similarity only considers feature membership and ignores ranking positions, it measures agreement in feature selection rather than agreement in terms of rank order. However, the ranking behaviour can be observed as we are comparing Jaccard similarities at different depths.

Rank-Biased Overlap. Similarly, RBO [23] is calculated between two ranked feature lists, and measures the similarity between rankings, but contrary to Jaccard similarity, it assigns more weight to higher-ranked items and gradually decreases the influence of lower-ranked positions. Formally, Rank-Biased Overlap can be defined as in Equation 3 (taken from [23]):

$$\mathrm{RBO}(S, T, p) = (1 - p) \sum_{d=1}^{\infty} p^{d-1} A_d \tag{3}$$

Here, S and T are two infinite rankings. A persistence parameter p between 0 and 1 controls the top-weighted emphasis where values closer to 1 consider deeper ranks more strongly. At each depth d, the overlap between the first d items of both rankings is computed and weighted by p^{d-1}. A_d represents the overlap between S and T at d, calculated as the proportion of common items among the top d positions.

All experiments use $p = 0.9$, and $d = 10$. The persistence parameter is set to give greater weight to highly ranked features as those are the top feature-wise contributions are more important for understanding an overall anomaly score. The depth is chosen according to the examined depths in the Jaccard similarity analysis.

The resulting score ranges from 0 to 1, where 0 indicates no overlap between the rankings and 1 indicates identical rankings. As the original RBO paper [23] does not suggest any interpretations for the scores, this thesis sets its own thresholds, where values below 0.2 indicate weak agreement, values between 0.2 and 0.5 indicate moderate agreement, and values above 0.5 indicate strong agreement.

Spearman's Rank Correlation Coefficient. Spearman's rank correlation coefficient (ρ) [22] is calculated between two feature rankings and measures the strength and direction of the association between their ranking positions. Features that are absent from one ranking are assigned a rank of $k + 1$, where k is the maximum ranking depth considered in the comparison. The resulting coefficient ranges from -1 to 1, where a value of 1 indicates identical ranking order, a value of 0 indicates no association between the rankings, and a value of -1 indicates completely reversed ranking order. In this thesis, absolute values below 0.3 indicate weak association, values between 0.3 and 0.7 indicate moderate association, and values above 0.7 indicate strong association.

As Spearman's correlation is a non-parametric measure, it is not required to assume that the rankings follow a normal distribution. The null hypothesis ($H_0 : \rho = 0$) states that there is no association between the rankings, while the alternative hypothesis ($H_a : \rho \neq 0$) states that an association exists between the rankings. To evaluate statistical significance, a significance level of $\alpha = 0.05$ is used. Similarly to the other evaluation methods in this subsection, Spearman's rank correlation coefficient is computed at a ranking depth of $k = 10$ to test for a statistically significant association between the rankings.

5 Results and Discussion

5.1 RQ1: Evaluating TCCM's Self-Explanations

This section presents and discusses the results of the first research question: *To what extent do TCCM's intrinsic feature-wise importance scores identify the same features that contribute to an anomaly compared to post-hoc explanation methods applied on the same TCCM model?*

5.1.1 Results

Anomaly Detection Performance. Table 2 shows an overview of TCCM's AUROC with each dataset's characteristics. Overall, TCCM's predictive performance varies across the evaluated datasets. It achieves the highest AUROC on the *UNSW-NB15* (0.9967) and *cover* datasets (0.9767), in which cases the model separates normal instances and anomalies almost perfectly. The *magic.gamma* dataset achieves a somewhat lower AUROC (0.8080), while *celeba* has the weakest anomaly detection performance (0.6446).

This demonstrates that TCCM is the most effective on the cybersecurity dataset that has a significantly larger sample size, and struggles the most with the facial attributes dataset that contains exclusively binary attributes. Therefore, the distinction between normal samples and anomalies with solely binary attributes seem to be less separable in the learnt contraction space.

Dataset	Features	Anomaly	Explained	TCCM AUROC
celeba	39	4 547	4 547	0.6446
cover	10	2 747	2 747	0.9767
magic.gamma	10	6 688	6 688	0.8080
UNSW-NB15	290	22 215	22 215	0.9967

Table 2: Summary of the dataset characteristics and TCCM's classification performance. For each dataset, the table shows the number of features, the number of anomalous samples, the number of anomalies that are explained by SHAP (applied on TCCM) and TCCM, and the AUROC score achieved by TCCM.

Ranking Agreement. In terms of the ranking agreement, TCCM's intrinsic feature-wise importance scores and SHAP values significantly differ between datasets. For an overview of the results of the explainability evaluation, see Table 3. As the *cover* and the *magic.gamma* datasets only contain 10 features in total, it is to be expected that they both achieve complete overlap at the top 10 level, with a Jaccard similarity of 1. The RBO values of these datasets are also the highest among the results, with *cover* having a score of 0.4202 and *magic.gamma* a score of 0.4334. These results can occur to be higher than the others due to their lower dimensionality (10), as all of their features are going to be included in the top ten features.

On the other hand, the *celeba* dataset shows the lowest agreement between TCCM and SHAP, which is reflected consistently across every evaluation metric. The RBO score of 0.0781 shows a very weak agreement, and the Jaccard similarity highlights a very limited overlap at the deeper levels,

e.g., the top 10 level score of 2/18. This indicates that the two explanation methods prioritized very different features and ranking orders.

Lastly, the *UNSW-NB15* dataset demonstrates intermediate behaviour. While the overlap at shallow ranking depths remains low (no overlap at the first level, and only one feature overlap at top 3), the agreement improves at deeper levels, where they agree on 3 features at the top 5 level, and reaching a Jaccard at 10 score of 5/15 and an RBO value of 0.2443. This suggests moderate similarity between the network traffic features found by TCCM and SHAP.

Dataset	J-1	J-3	J-5	J-10	RBO	ρ	p-value
celeba	0/2	0/6	1/9	2/18	0.0781	-0.6249	0.0056
cover	1/1	1/5	2/8	10/10	0.4202	0.2364	0.5109
magic.gamma	0/2	2/4	3/7	10/10	0.4334	0.6606	0.0376
UNSW-NB15	0/2	1/5	3/7	5/15	0.2443	-0.1296	0.6452

Table 3: Explainability comparison results for each dataset. This table shows the Jaccard similarity between the top 1, 3, 5 and 10 features in the feature rankings of TCCM and SHAP on TCCM (denoted as J-1, J-3, J-5, and J-10). It also reports the Rank-Biased Overlap (RBO) score at depth 10, Spearman's rank correlation coefficient ρ, and the corresponding p-value at significance level 0.05 for statistical significance.

Statistical Testing. Spearman's rank correlation analysis further highlighted differences between the datasets. The strongest positive correlation was observed for the *magic.gamma* dataset ($\rho = 0.6606$, $p = 0.0376$), which indicates a statistically significant moderate association between the TCCM and SHAP feature rankings. The *cover* dataset showed a weak positive correlation ($\rho = 0.2364$, $p = 0.5109$), while *UNSW-NB15* ($\rho = -0.1296$, $p = 0.6452$) and *celeba* ($\rho = -0.6249$, $p = 0.0056$) showed negative correlations. Among the evaluated datasets, only *magic.gamma* and *celeba* produced statistically significant correlations at the 0.05 significance level.

The results of the *magic.gamma* dataset indicate that features that TCCM considers important also tend to receive higher SHAP rankings. On the other hand, the *celeba* dataset exhibits a significant negative correlation, which suggests that features ranked highly by one explanation method often appear at lower positions in the other ranking. The remaining datasets do not produce statistically significant correlations, which can mean that the observed similarities may be limited to specific features.

Ordering of Features. Figures 1 and 2 show comparisons of the top 10 features contributing to an anomaly score. The order of features in the rankings are the most aligned between TCCM and SHAP in the *magic.gamma* dataset. Both methods identify the same top two features, Width and Size, however, the order of these features is flipped between the methods. Other features, such as Length, Alpha (the angle of major axis with vector to origin), and Conc (the ratio of the sum of the two highest pixels over Size) are also all selected in the top 6 places, with differing ranking positions. These results indicate that both methods can identify the most prominent features contributing to a hadronic shower relatively similarly, while they also place the least influential features in very similar positions.

As for the *cover* dataset, it shows substantial consistency between the two explanation approaches. Both methods identify Elevation as the dominant feature for determining a forest cover type, while some of the other attributes (Horizontal Distance to Roadways, Vertical Distance to Hydrology, and Hillshade at Noon) also appear in similar positions in both rankings. The shape of Figure 1b shows that TCCM assigns more balanced importance scores across the features opposed to the SHAP values of TCCM anomalies, where the first three features appear more prominently than the lower ranking features. These results indicate that in determining a forest cover type, Elevation seems to be an unequivocally defining feature that both methods recognise.

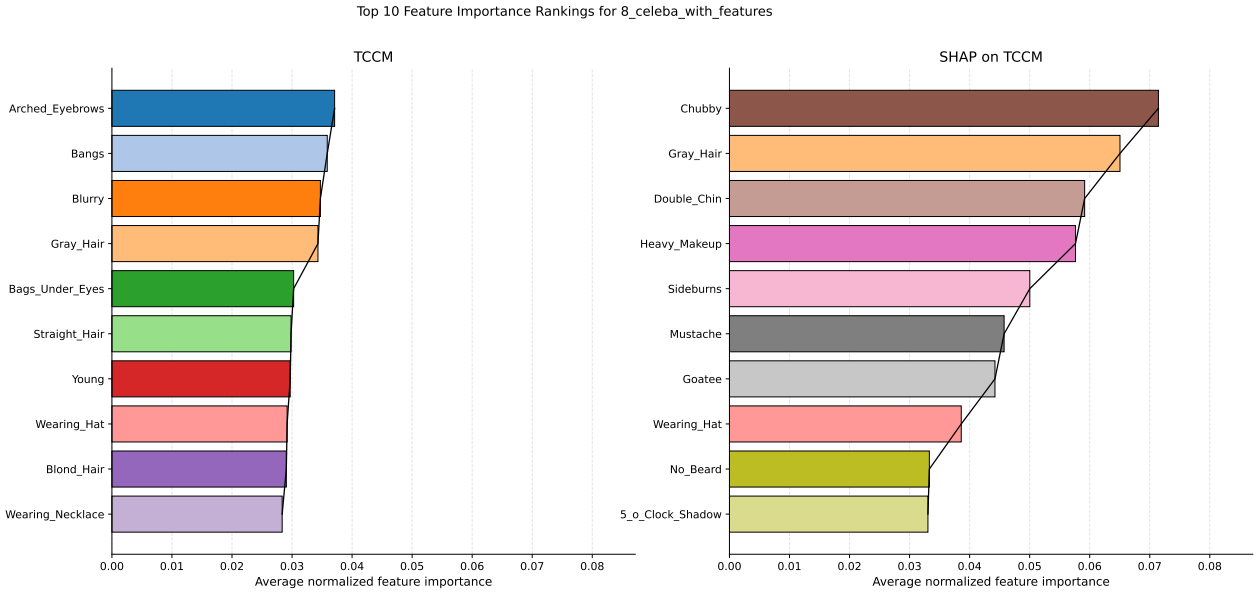
In contrast, the *celeba* dataset reveals clear differences between the intrinsic and post-hoc explanations. TCCM's top attributes in determining whether someone is bald are Arched Eyebrows, Bangs, and Blurry (picture), while SHAP assigns higher importance to features Chubby, Gray Hair, and Double Chin. Wearing Hat is ranked at the same position in both rankings, while Gray Hair is ranked high on both lists. This highlights how TCCM's internal structure may be struggling with this dataset in particular as Arched Eyebrows or the blurriness of the picture do not seem to be meaningful explanations for baldness.

The *UNSW-NB15* dataset also demonstrates notable differences between the two explanation methods. TCCM focuses on time-to-live (TTL) and IP-addresses, whereas SHAP strongly prioritises specific IP-address features. Out of the 290 features, the two methods share five features in their rankings, however, those are taking up different positions. The two methods agree on the second feature, which is a specific source IP address, and they place source to destination TTL very similarly. However, SHAP places a certain destination IP first that does not even appear in the TCCM ranking. Overall, this indicates that the two explanation methods identified a small subset of common important features, but varied in how they ranked those features and in which additional features they considered important.

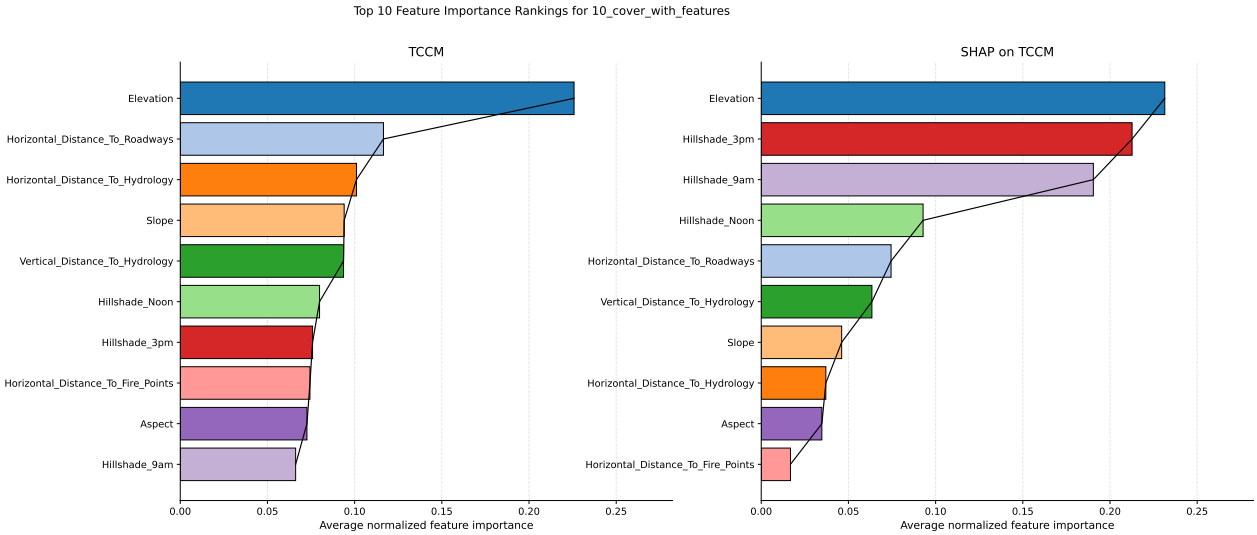
5.1.2 Discussion

The results show that the agreement between TCCM's intrinsic feature-wise importance scores and SHAP post-hoc explanations of TCCM scores is dependent on dataset characteristics. Additionally, while moderate agreement can be observed on most datasets, the rankings differ both in selected features and in their ordering, which suggests that TCCM's intrinsic importance scores and SHAP's post-hoc feature attributions catch different aspects when explaining the model's anomaly scores.

The highest Jaccard similarities are observed on the *cover* and *magic.gamma* datasets, however, both datasets contain only 10 features, which increases the probability of overlap between the top ranked features. Therefore, the Jaccard scores across datasets with very different feature dimensionalities are not fully comparable as the metric becomes less discriminative for low-dimensional datasets. On the other hand, the RBO metric is more informative of ranking consistency because it also considers the ordering of the shared features and assigned greater importance to higher positions. The *cover* and *magic.gamma* datasets also achieve the highest RBO scores with moderate agreement, not only in the selected features but also in their ranking positions. The *celeba* dataset produces the weakest agreement between the explanation methods. Consequently, SHAP explanations of the TCCM anomaly scores highlight different facial attributes than TCCM's intrinsic interpretations. One possible reason for this behaviour is that the dataset consists exclusively of binary facial attributes, which may be more difficult to separate in the learnt contraction space of TCCM. The

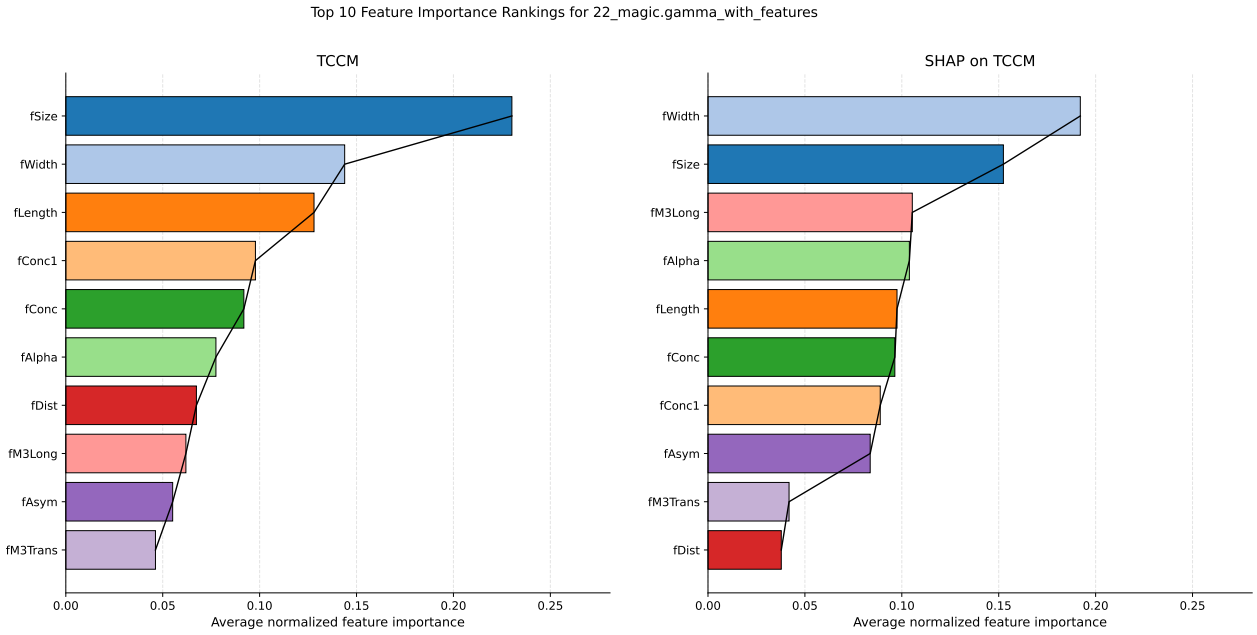


(a) *celeba*

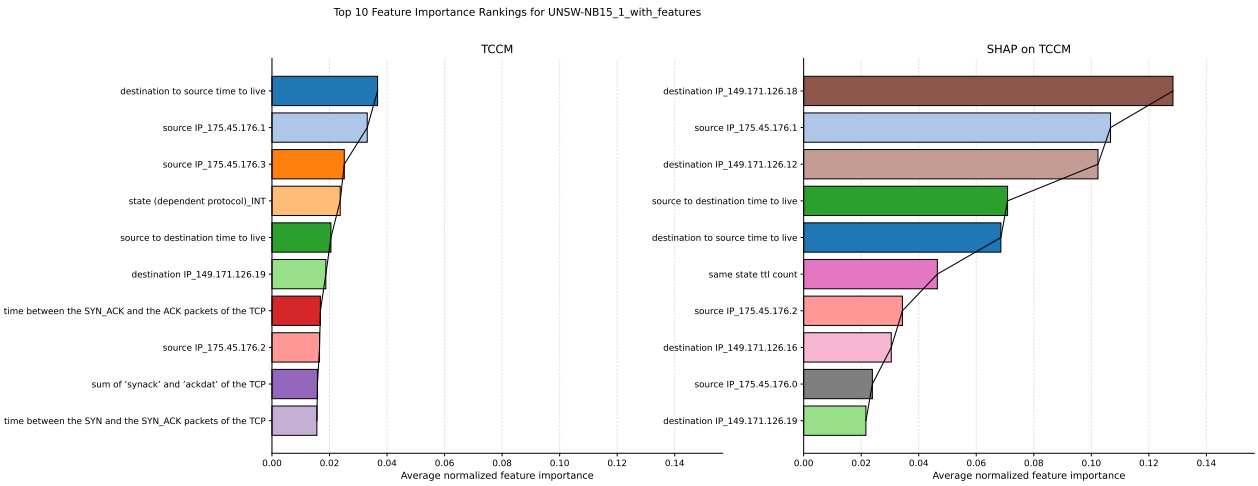


(b) *cover*

Figure 1: Feature ranking comparisons for the first research question on the *celeba* and *cover* datasets. The left column shows TCCM's intrinsic feature importance rankings, while the right column shows SHAP explanations.



(a) *magic.gamma*



(b) *UNSW-NB15*

Figure 2: Feature ranking comparisons for the first research question on the *magic.gamma* and *UNSW-NB15* datasets. The left column shows TCCM's intrinsic feature importance rankings, while the right column shows SHAP explanations.

UNSW-NB15 dataset demonstrates intermediate behaviour. Despite the high dimensionality of the dataset (290 features), the two methods still share several features within their rankings. The strong emphasis on IP-address features in the SHAP explanations may suggest that the post-hoc explanations caught statistical patterns that are less prevalent in TCCM's intrinsic representations. Therefore, it is possible that SHAP and TCCM capture different aspects of anomalous behaviour.

In conclusion, the results suggest that SHAP explanations of TCCM anomaly scores do not necessarily reproduce the model's intrinsic feature-wise scoring. This difference may be related to the different internal mechanisms of the two approaches. SHAP explains the influence of individual features on the final anomaly score to changes in the input features, whereas TCCM's intrinsic importance scores are derived directly from the model's learnt contraction field and residuals. Consequently, the explanation scores produced by the two methods may not be fully comparable. Nevertheless, the moderate overlap observed on several datasets indicates that TCCM's intrinsic importance scores and SHAP explanations of TCCM anomalies still highlight some shared patterns in anomalies.

5.2 RQ2: Comparing Intrinsic and Post-hoc Explanations

This section presents and discusses the results of the second research question: *How does TCCM's intrinsic interpretability compare to the explanations generated by post-hoc explanation methods applied on a different deep learning model?*

5.2.1 Results

Anomaly Detection Performance. Table 4 shows an overview of TCCM's and DTE's AUROC with each dataset's characteristics. The anomaly detection performance of both methods varies across the evaluated datasets. TCCM achieves the highest AUROC on the *UNSW-NB15* dataset (0.9973), while DTE achieves a slightly lower AUROC of 0.9823 on the same dataset. The two methods show comparable performance on the *celeba* dataset, where TCCM achieves an AUROC of 0.6494 and DTE achieves 0.6624, and the *cover* dataset (TCCM: 0.9878, DTE: 0.9892). On the *magic.gamma* dataset, DTE obtains an AUROC of 0.8410 compared to TCCM's AUROC of 0.8080.

The results of the second research question demonstrate that DTE generally achieves slightly stronger anomaly detection performance than TCCM across almost all evaluated datasets. The largest performance difference can be observed on the *magic.gamma* dataset, while the performance of the two methods remains relatively similar on *celeba*, *cover*, and *UNSW-NB15*.

Dataset	Features	Anomaly	Explained	TCCM AUROC	DTE AUROC
celeba	39	4 547	100	0.6494	0.6624
cover	10	2 747	100	0.9878	0.9892
magic.gamma	10	6 688	100	0.8080	0.8410
UNSW-NB15	290	22 215	100	0.9973	0.9823

Table 4: Summary of the dataset characteristics and the classification performance of TCCM and DTE. For each dataset, the table shows the number of features, the number of anomalous samples, the number of anomalies that are explained by SHAP (applied on DTE) and TCCM, and the AUROC score achieved by TCCM and DTE.

Dataset	J-1	J-3	J-5	J-10	RBO	ρ	p-value
celeba	0/2	0/6	1/9	3/17	0.0713	-0.6947	0.0020
cover	1/1	1/5	3/7	10/10	0.4440	0.3455	0.3282
magic.gamma	0/2	2/4	2/8	10/10	0.3345	0.0788	0.8287
UNSW-NB15	1/1	2/4	3/7	6/14	0.4925	0.4809	0.0817

Table 5: Explainability comparison results for each dataset. This table shows the Jaccard similarity between the top 1, 3, 5 and 10 features in the feature rankings of TCCM and SHAP on DTE (denoted as J-1, J-3, J-5, and J-10). It also reports the Rank-Biased Overlap (RBO) score at depth 10, Spearman's rank correlation coefficient ρ, and the corresponding p-value at significance level 0.05 for statistical significance.

Ranking Agreement. In terms of the ranking agreement (see Table 5), TCCM's intrinsic feature-wise importance scores and SHAP explanations of DTE differ considerably between datasets. Similar to the first research question, the *cover* and *magic.gamma* datasets contain only 10 features, therefore complete overlap at the top 10 level was expected. However, the ordering of the shared features significantly varies, which is reflected by the moderate RBO values.

The *cover* dataset achieves moderate agreement, with complete overlap at the top 1 level, a Jaccard similarity of 3/7 at the top 5 level, complete overlap at the top 10 level, and an RBO value of 0.4440. The *magic.gamma* dataset shows somewhat weaker agreement, achieving no overlap at the top 1 level, a Jaccard similarity of 2/8 at the top 5 level, and an RBO value of 0.3345.

The *celeba* dataset demonstrates the lowest agreement between TCCM and SHAP on DTE. There is no overlap at the top 1 and top 3 levels, no shared features at the top 5 level, and only limited overlap at deeper ranking levels with a Jaccard at 10 score of 3/17 and an RBO value of 0.0713. The low overlap and very small RBO value indicate that the two methods usually focused on different attributes when identifying anomalous instances, just as in RQ1.

The *UNSW-NB15* dataset achieves the strongest ranking agreement among the evaluated datasets. The two methods share the highest overlap at the top-ranked feature, resulting in a Jaccard similarity of 1. The agreement remains relatively strong at deeper ranking depths, reaching 3/7 at the top 5 level and 6/14 at the top 10 level. The dataset also achieves the highest RBO value (0.4925), indicating moderate agreement in both feature selection and ordering.

Statistical Testing. The Spearman's rank correlation coefficient is additionally measured to evaluate whether the ordering of the rankings is associated. The strongest positive correlations are observed on the *UNSW-NB15* ($\rho = 0.4809$, $p = 0.0817$) and the *cover* ($\rho = 0.3455$, $p = 0.3282$) datasets, while the *magic.gamma* shows almost no correlation between the rankings ($\rho = 0.0788$, $p = 0.8287$). The *celeba* dataset produces a statistically significant negative correlation ($\rho = -0.6947$, $p = 0.0020$), which highlights substantial disagreement in the ordering of the features between TCCM and SHAP on DTE. However, none of the positive correlations reach statistical significance at the 0.05 level.

These correlation results provide a more detailed view of ranking agreement by considering the ordering of the features. The *celeba* dataset producing a statistically significant negative correlation suggests that facial features receiving high ranks from one explanation method fre-

quently appear at lower ranks in the other ranking. Regarding the *cover* and the *UNSW-NB15* datasets, they have moderate positive correlations, which points to some consistency in ranking order, although the correlations are not statistically significant. The *magic.gamma* dataset shows no correlation, which indicates that the two methods frequently arrange features in very different orders.

Ordering of Features. The feature rankings in Figures 3 and 4 further illustrate these differences. The *cover* dataset shows the strongest consistency between the two explanation approaches. Both methods identify Elevation as the most influential feature and also rank some of the other attributes in similar positions, including Horizontal Distance to Fire Points and Slope. Although the ordering of the features differs, the overall structure of the rankings remains similar. While both methods assign a dominant role to Elevation, the remaining features receive substantially lower importance scores, particularly from SHAP.

For the *magic.gamma* dataset, the rankings show partial agreement. Both methods assign high importance to Width and Length, although their relative positions slightly differ. A notable difference is observed for Size, which was the most important feature according to TCCM but appears at the bottom of the SHAP ranking. Similarly, Alpha is ranked sixth by TCCM but second by SHAP on DTE. This shows that the two methods moderately focus on the same set of hadronic shower characteristics, the ranking positions and importance magnitudes differ considerably.

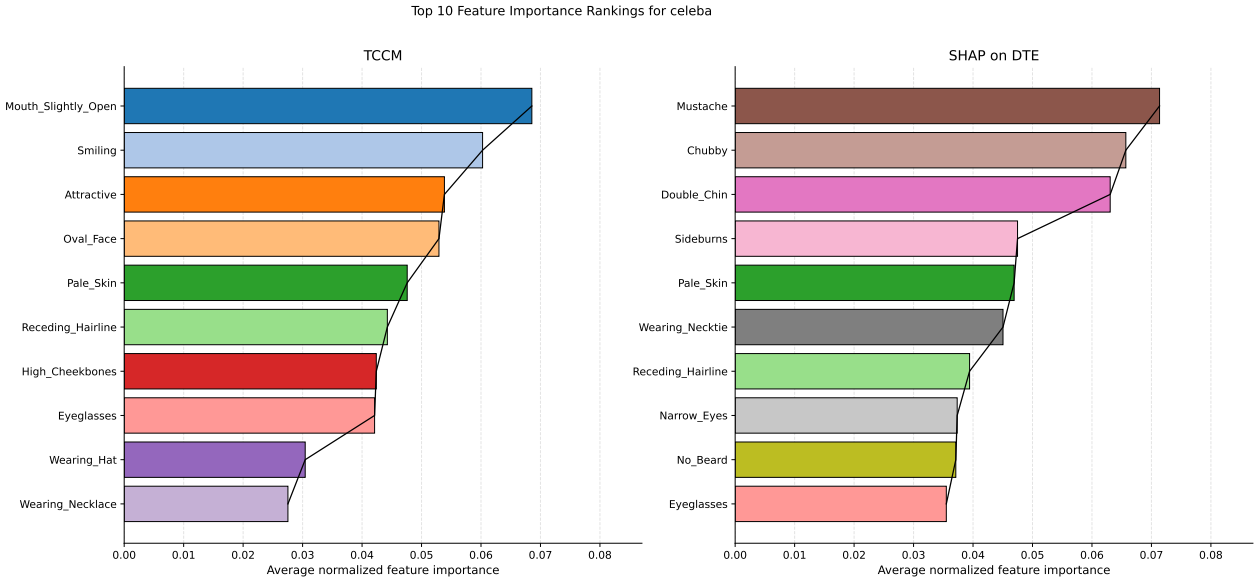
The rankings of the *UNSW-NB15* dataset show that both methods emphasize several IP-address attributes. Although the exact order differs, the rankings share multiple features throughout the top 10 positions. On this dataset, TCCM's intrinsic explanations and SHAP explanations of DTE achieve greater consistency than on the other datasets.

The *celeba* dataset demonstrate the clearest disagreement between the explanation methods. TCCM assigns the highest importance to facial attributes such as Mouth Slightly Open, Smiling, and Attractive, while SHAP on DTE emphasized features including Double Chin, Chubby, and Mustache. Some features, such as Pale Skin, Eyeglasses, and Receding Hairline, appear in both rankings in very similar positions. Furthermore, the dominant features highlighted by SHAP on DTE are absent from TCCM's highest-ranked features, which indicates that the two approaches rely on different aspects of the data when they produce their explanations.

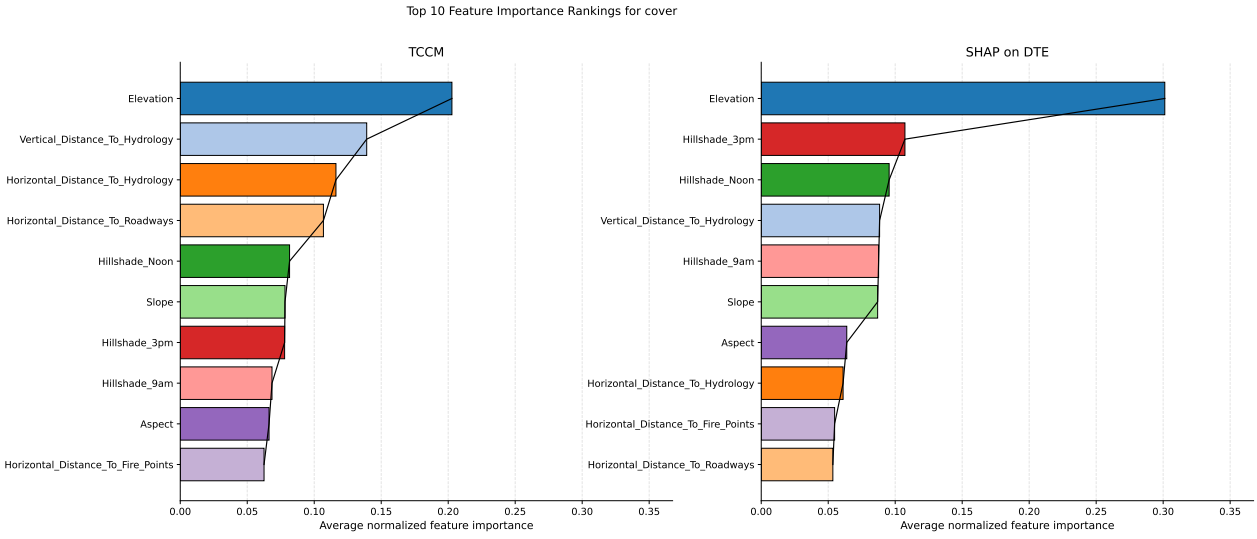
5.2.2 Discussion

Similarly to the first research question, the agreement between TCCM's intrinsic feature-wise importance scores and SHAP explanations is highly dependent on dataset characteristics. However, an important distinction compared to the first research question is that SHAP explanations are computed for a fixed subset of 100 detected anomalies in order to reduce computational cost. Consequently, the resulting feature rankings represent a sample of the anomalous instances rather than the complete anomaly population. Although this approach reduces the training time for DTE, some variability in the rankings may be caused by the selected subset. This would explain why the agreement between TCCM and SHAP on DTE is generally slightly higher than the agreement previously observed between TCCM and SHAP on TCCM.

In terms of the Jaccard similarities, the highest agreement can be observed on the *cover* and *UNSW-NB15* datasets. However, to highlight again, the *cover* and *magic.gamma* datasets only contain 10 features, therefore complete overlap at the top 10 level is inevitable and the probability

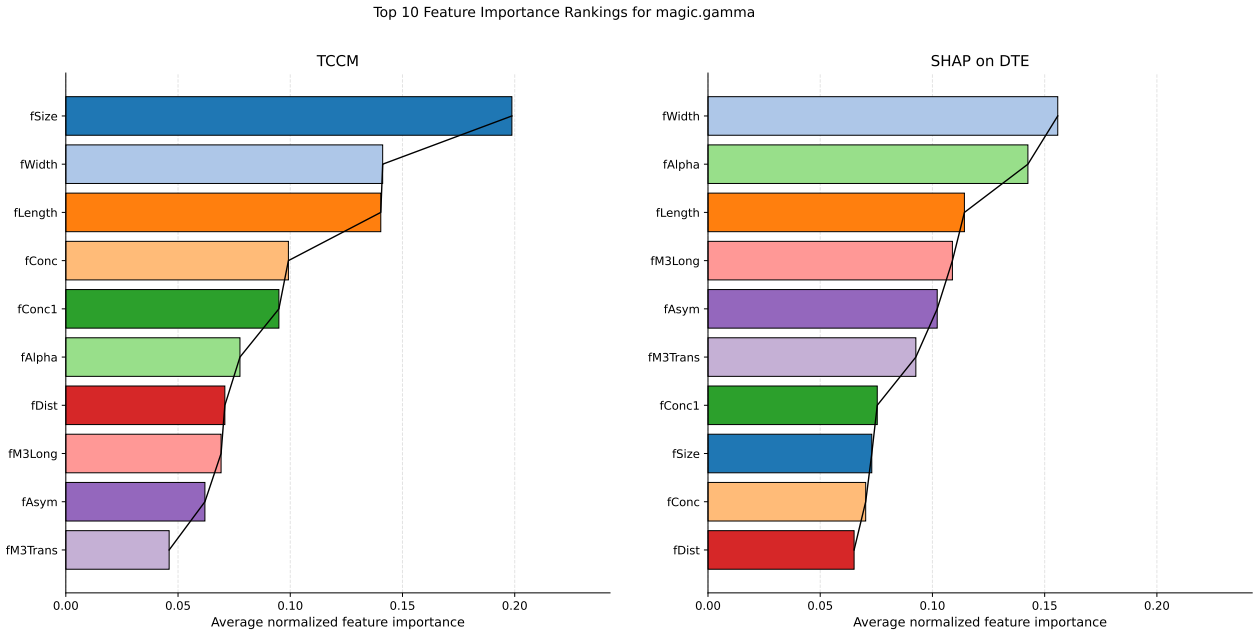


(a) *celeba*

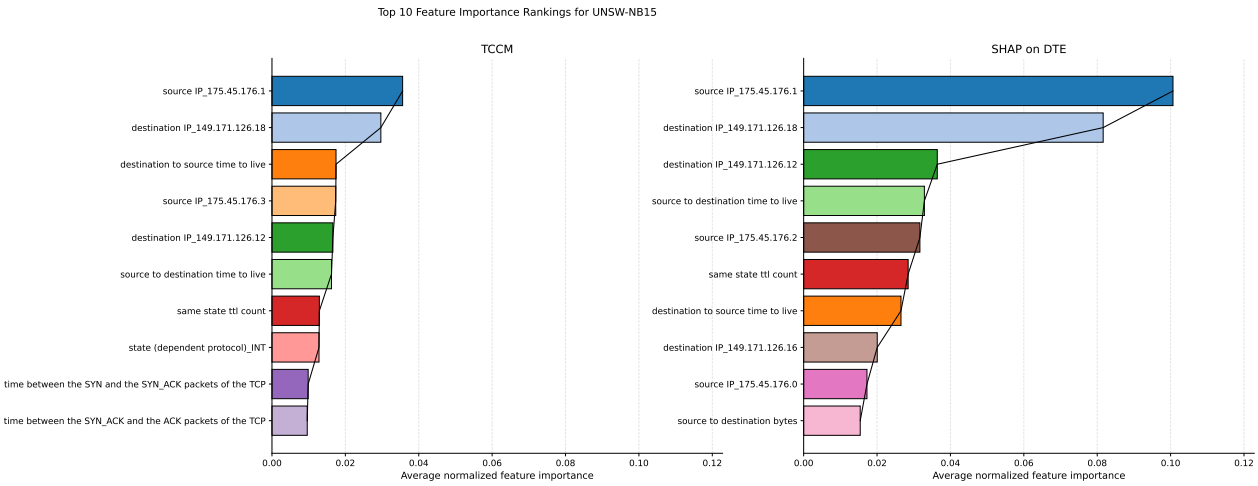


(b) *cover*

Figure 3: Feature ranking comparisons for the second research question on the *celeba* and *cover* datasets. The left rankings show TCCM's intrinsic feature importance scores, while the right rankings show SHAP explanations of DTE anomaly scores.



(a) *magic.gamma*



(b) *UNSW-NB15*

Figure 4: Feature ranking comparisons for the second research question on the *magic.gamma* and *UNSW-NB15* datasets. The left rankings show TCCM's intrinsic feature importance scores, while the right rankings show SHAP explanations of DTE anomaly scores.

of feature overlap at shallower depths is increased. Consequently, the Jaccard scores across datasets with very different feature dimensionalities are not fully comparable as the metric becomes less discriminative for low-dimensional datasets.

Similar to the first research question, the RBO metric is more informative of ranking consistency. While *cover* achieves an RBO score of 0.4440, which is very close to the value observed in the first research question (0.4202), the RBO score of *magic.gamma* decreases from 0.4334 to 0.3345. This indicates that although the two methods select some of the same features, their ranking positions differ more when SHAP is applied to DTE rather than TCCM. The *celeba* dataset repeatedly produces the weakest agreement between the explanation methods. Identically to the first research question, there is no overlap at the top 1 and top 3 levels, while the RBO value remains very low. Furthermore, the dominant facial attributes identified by SHAP on DTE differ a lot from those highlighted by TCCM's intrinsic explanations. This is consistent with the findings of the first research question and suggests that the binary facial attributes may be represented differently by the two explanation approaches. These results indicate that TCCM and DTE rely on different subsets of facial characteristics when distinguishing normal instances from anomalies. Overall, the *cover* dataset demonstrates the second strongest consistency between the methods. The agreement levels are comparable to those observed in the first research question, suggesting that the importance of these forest cover type attributes is relatively stable regardless of whether SHAP is applied to TCCM or DTE anomaly scores. The *UNSW-NB15* dataset achieves the highest RBO score among all datasets. This suggests that TCCM's intrinsic explanations and SHAP explanations of DTE showed relatively strong agreement regarding the most influential features. For example, several shared IP-address features appeared near the top of both rankings. Because RBO places greater emphasis on highly ranked features, this agreement at shallow ranking depths contributed substantially to the overall similarity score.

In summary, the results suggest that SHAP explanations of DTE anomaly scores do not necessarily reproduce the feature-wise importance scores produced intrinsically by TCCM. This difference is expected because the two methods rely on fundamentally different anomaly detection mechanisms. TCCM derives its intrinsic importance scores directly from the learnt contraction field and residual vectors, while DTE is a density-based method and SHAP explains the contribution of individual features to those scores. Consequently, the explanation scores produced by the two approaches are not directly comparable. Nevertheless, the moderate agreement observed on several datasets indicates that both methods can still identify some common patterns associated with anomalous instances.

5.3 RQ3: Accuracy-Interpretability Trade-off

This section presents and discusses the results of the third research question: *Does TCCM achieve a better balance between accuracy and interpretability compared to classical interpretable machine learning approaches?*

5.3.1 Decision Tree Results

Anomaly Detection Performance. In terms of predictive performance, the Decision Tree achieves higher AUROC values than TCCM on all evaluated datasets (see Table 6). The largest differences are shown on the *magic.gamma* dataset, where the Decision Tree achieves an AUROC

of 0.8576 compared to TCCM's 0.5128, and on the *celeba* dataset, where the Decision Tree reaches 0.9087 compared to 0.7522 for TCCM. On the *cover* and *UNSW-NB15* datasets, both methods achieve almost perfect anomaly detection performance, with the Decision Tree slightly outperforming TCCM.

Dataset	Features	Anomaly	TCCM AUROC	Tree AUROC
celeba	39	4 547	0.7522	0.9087
cover	10	2 747	0.9228	0.9980
magic.gamma	10	6 688	0.5128	0.8576
UNSW-NB15	290	22 215	0.9978	0.9988

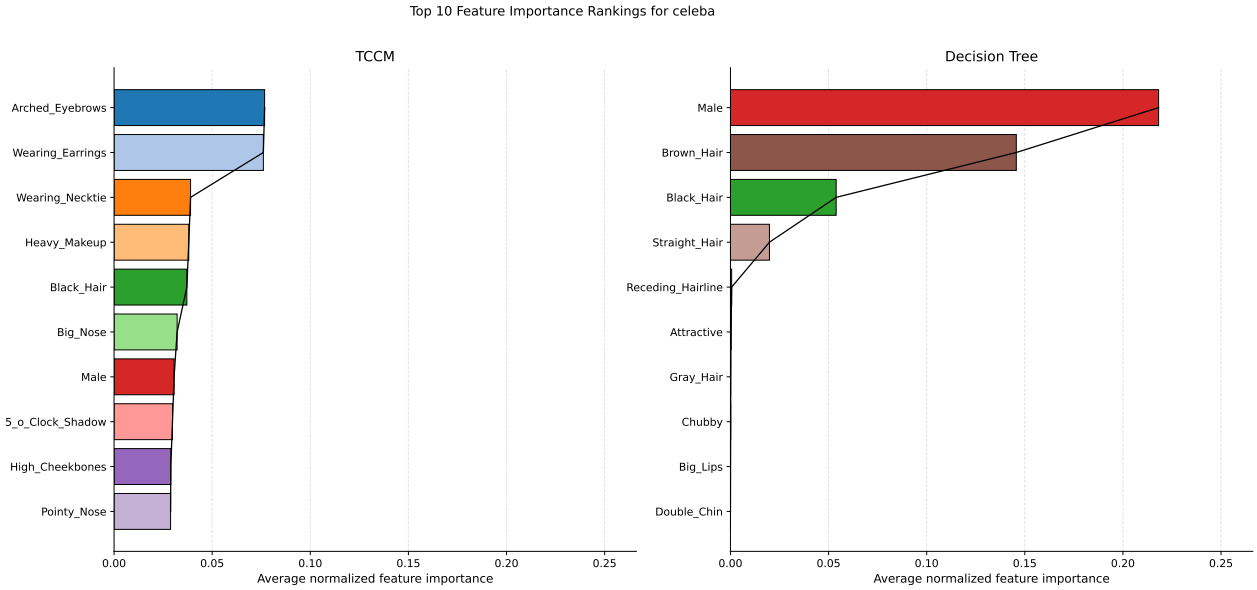
Table 6: Summary of the dataset characteristics and the classification performance of TCCM and Decision Tree. For each dataset, the table shows the number of features, the number of anomalous samples, and the AUROC score achieved by TCCM and Decision Tree.

Dataset	J-1	J-3	J-5	J-10	RBO	ρ	p-value
celeba	0/2	0/6	1/9	2/18	0.0674	-0.5412	0.0204
cover	0/2	0/6	2/8	10/10	0.2224	-0.3576	0.3104
magic.gamma	0/2	1/5	2/8	10/10	0.2723	-0.0788	0.8287
UNSW-NB15	0/2	0/6	0/10	1/19	0.0321	-0.7020	0.0008

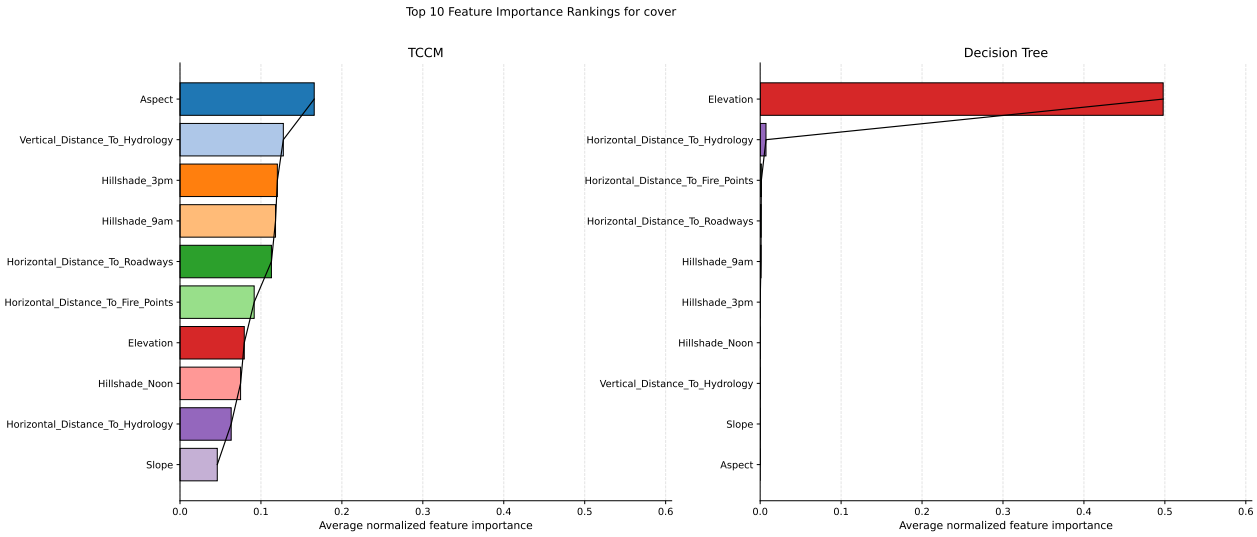
Table 7: Explainability comparison results for each dataset. This table shows the Jaccard similarity between the top 1, 3, 5 and 10 features in the feature rankings of TCCM and Decision Tree (denoted as J-1, J-3, J-5, and J-10). It also reports the Rank-Biased Overlap (RBO) score at depth 10, Spearman's rank correlation coefficient ρ, and the corresponding p-value at significance level 0.05 for statistical significance.

Ranking Agreement. Despite the strong predictive performance of the Decision Tree, the agreement between its feature importance rankings and TCCM's intrinsic scores remains low as illustrated in Table 7. The *cover* dataset achieves an RBO score of 0.2224. Similarly, the *magic.gamma* dataset has a slightly higher RBO value of 0.2723. However, neither dataset shares any features at the top 1 level, indicating that the methods disagree on the most important feature despite agreeing on several lower-ranked features. The *celeba* dataset demonstrates weak agreement between the rankings. Only one feature overlaps at the top 5 level and two features at the top 10 level, resulting in an RBO score of 0.0674. The *UNSW-NB15* dataset shows the weakest agreement overall, with no shared features at the top 5 level, only one shared feature at the top 10 level, and an RBO score of 0.0321. These feature rankings reveal core differences between the two approaches.

Statistical Testing. The Spearman's rank correlation coefficient shows whether the ordering of the ranking is associated. For every dataset, we observe negative correlations, with the *UNSW-NB15* dataset showing the strongest association ($\rho = -0.7020$, $p = 0.0008$), and the *magic.gamma* showing the weakest ($\rho = -0.0788$, $p = 0.8287$). The *celeba* ($\rho = -0.5412$, $p = 0.0204$) and the *cover*

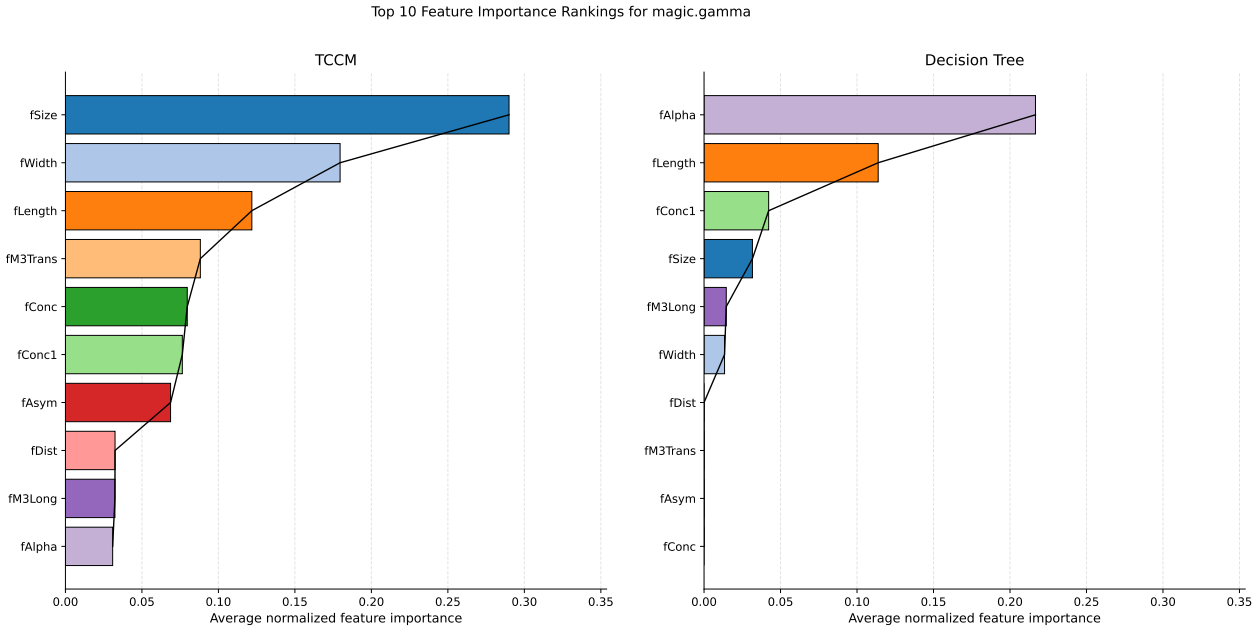


(a) *celeba*

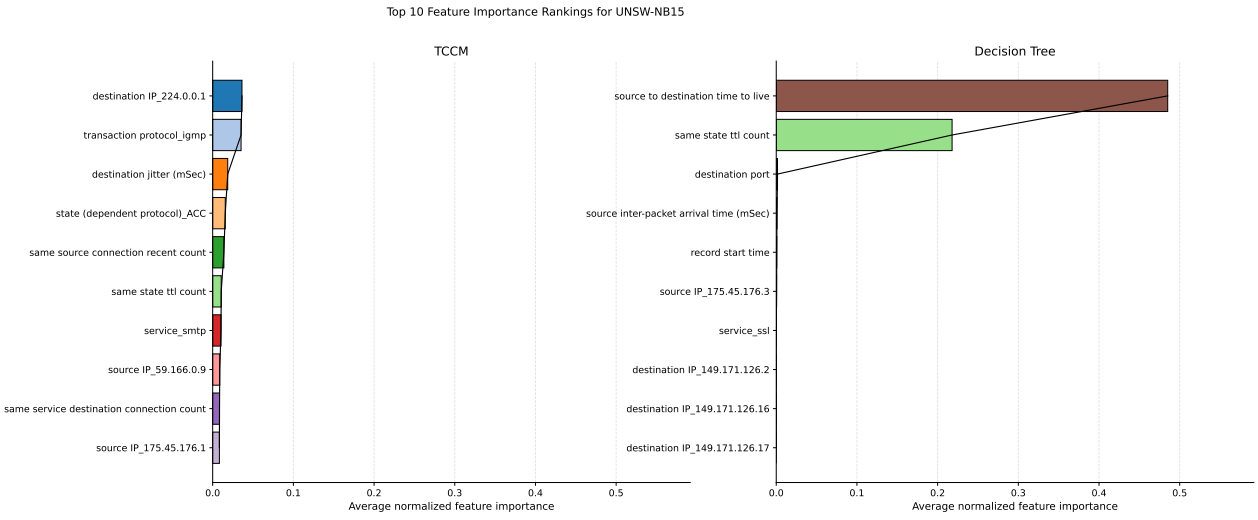


(b) *cover*

Figure 5: Feature ranking comparisons between TCCM and Decision Tree on the *celeba* and *cover* datasets. The left rankings show TCCM's intrinsic feature importance scores, while the right rankings show Decision Tree permutation feature importance scores.



(a) *magic.gamma*



(b) *UNSW-NB15*

Figure 6: Feature ranking comparisons between TCCM and Decision Tree on the *magic.gamma* and *UNSW-NB15* datasets. The left rankings show TCCM's intrinsic feature importance scores, while the right rankings show Decision Tree permutation feature importance scores.

($\rho = -0.3576$, $p = 0.3104$) datasets show correlation that is in-between the other two datasets' results. Only the *celeba* and the *UNSW-NB15* datasets achieve statistical significance at the 0.05 level.

These results suggest that the facial features in the *celeba* dataset and the network traffic features in the *UNSW-NB15* that received high ranks from one explanation method frequently appear at lower ranks in the other ranking. The *cover* dataset has moderate negative correlation, which points to some consistency in ranking order, although the correlation is not statistically significant. The *magic.gamma* dataset shows very weak correlation, which indicates that the two methods frequently arrange features in very different orders.

Ordering of Features. Figures 5 and 6 depict the feature rankings of TCCM and Decision Tree. On the *cover* dataset, TCCM distributes importance relatively evenly across features such as Aspect, Vertical Distance to Hydrology, and Hillshade attributes, whereas the Decision Tree only prioritises Elevation. A similar pattern occurs on *magic.gamma*, where TCCM assigns importance to a broader set of attributes, while the Decision Tree concentrates most importance on Alpha and Length. On *celeba*, TCCM highlights facial attributes such as Arched Eyebrows and Wearing Earrings, whereas the Decision Tree focuses on attribute Male and hair style attributes. Finally, on *UNSW-NB15*, TCCM emphasizes network protocol, timing, and IP address features, while the Decision Tree primarily relies on time-to-live features.

5.3.2 Isolation Forest Results

Anomaly Detection Performance. The predictive performance comparison between TCCM and Isolation Forest produces mixed results, as illustrated in Table 8. Isolation Forest outperforms TCCM on the *magic.gamma* dataset (0.7794 vs. 0.5128), while TCCM achieves higher AUROC values on the remaining datasets. The largest advantage for TCCM is observed on *cover*, where TCCM achieves 0.9228 compared to 0.8161 for Isolation Forest. On the *UNSW-NB15* dataset, both methods achieve very high performance, although TCCM remains slightly superior.

Dataset	Features	Anomaly	TCCM AUROC	IF AUROC
celeba	39	4 547	0.7522	0.7286
cover	10	2 747	0.9228	0.8161
magic.gamma	10	6 688	0.5128	0.7794
UNSW-NB15	290	22 215	0.9978	0.9585

Table 8: Summary of the dataset characteristics and the classification performance of TCCM and Isolation Forest. For each dataset, the table shows the number of features, the number of anomalous samples, and the AUROC score achieved by TCCM and Isolation Forest.

Ranking Agreement. In terms of ranking agreement seen in Table 9, Isolation Forest demonstrates higher consistency with TCCM than the Decision Tree on several datasets. The highest agreement is observed on *magic.gamma*, which achieves a Jaccard similarity of 3/7 at the top 5

level and the highest RBO score among all experiments (0.3574). The *cover* dataset again achieves complete overlap at the top 10 level due to its low dimensionality, although the RBO value remains moderate at 0.2074. The *celeba* dataset shows limited agreement, with only one shared feature at the top 5 level and two shared features at the top 10 level. Similarly, the *UNSW-NB15* dataset demonstrates very weak agreement. No features overlap at the top 5 level, and only two features appear in both top 10 rankings, resulting in an RBO value of 0.0359.

Dataset	J-1	J-3	J-5	J-10	RBO	ρ	p-value
celeba	0/2	1/5	1/9	2/18	0.1126	-0.5220	0.0263
cover	0/2	0/6	2/8	10/10	0.2074	-0.7212	0.0186
magic.gamma	0/2	2/4	3/7	10/10	0.3574	-0.1758	0.6272
UNSW-NB15	0/2	0/6	0/10	2/18	0.0359	-0.7006	0.0012

Table 9: Explainability comparison results for each dataset. This table shows the Jaccard similarity between the top 1, 3, 5 and 10 features in the feature rankings of TCCM and Isolation Forest (denoted as J-1, J-3, J-5, and J-10). It also reports the Rank-Biased Overlap (RBO) score at depth 10, Spearman's rank correlation coefficient ρ, and the corresponding p-value at significance level 0.05 for statistical significance.

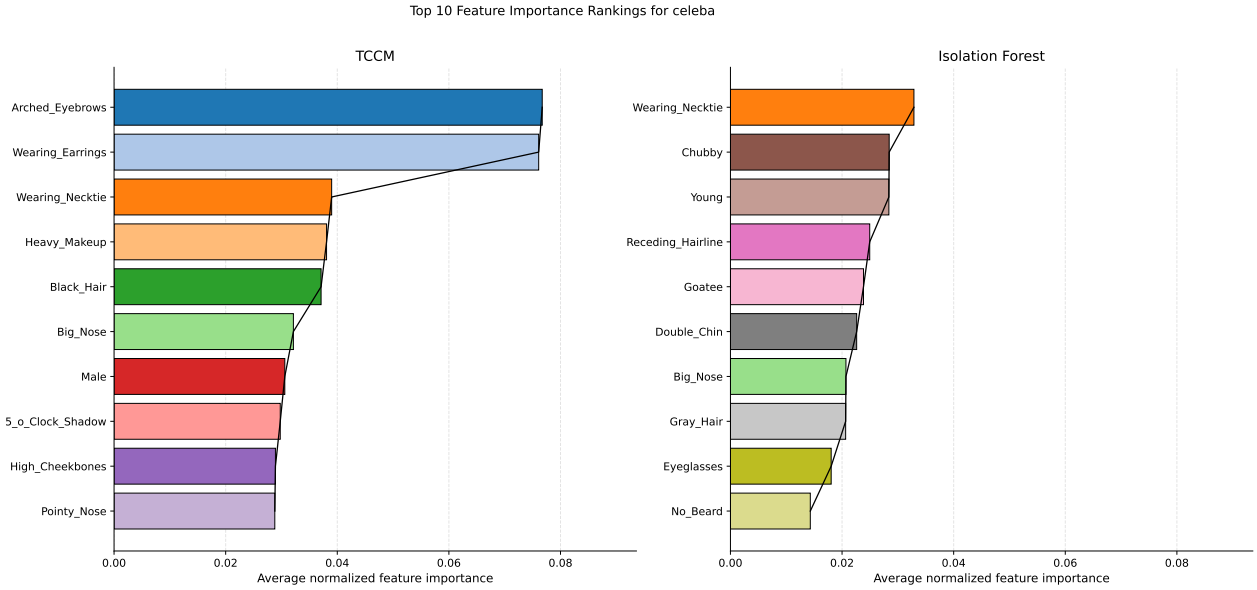
Statistical Testing. Similarly to the comparison with Decision Tree, the Spearman's rank correlation analyses all show negative correlations. However, almost all rankings reach statistical significance at the 0.05 level. The strongest correlation can be observed for the *cover* dataset ($\rho = -0.7212$, $p = 0.0186$), which indicates a statistically significant strong association between the TCCM and Isolation Forest feature rankings. The *UNSW-NB15* ($\rho = -0.7006$, $p = 0.0012$) dataset shows a similarly strong correlation that is statistically significant. The *celeba* ($\rho = -0.5220$, $p = 0.0263$) dataset shows a statistically significant moderate association, while the weakest association was produced by the *magic.gamma* dataset ($\rho = -0.1758$, $p = 0.6272$), although this is not statistically significant.

Compared with the previous experiments, these tests produced the highest number of statistically significant associations. Consequently, TCCM and Isolation Forest almost always rank features in a very different order. As the *magic.gamma* dataset did not produce a statistically significant correlation, the differences observed for these rankings may be caused by certain features.

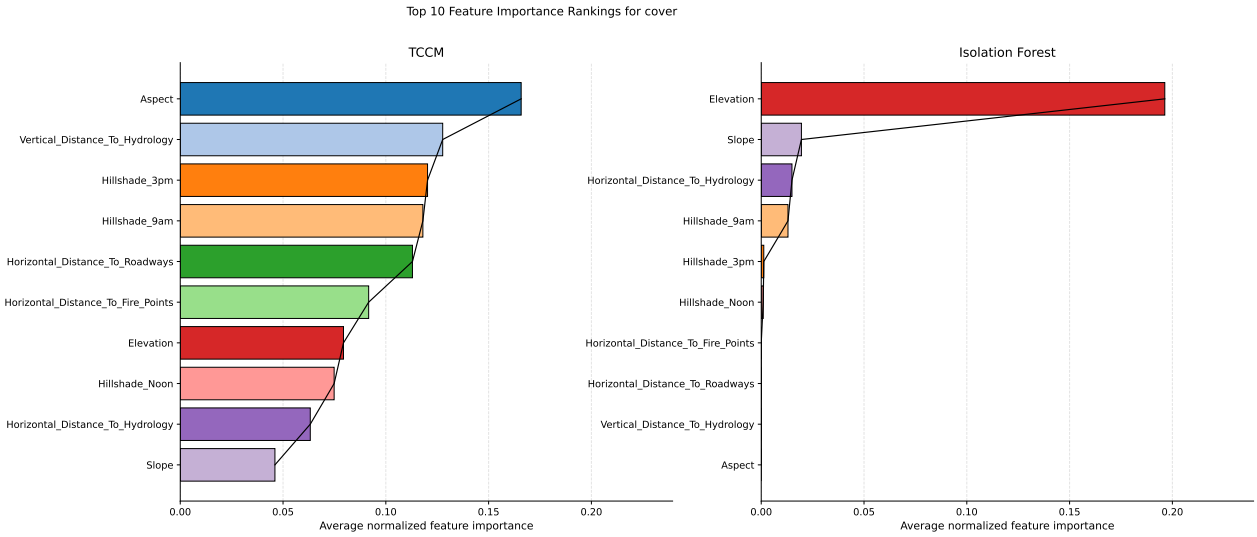
Ordering of Features. The ranking visualizations on Figures 7 and 8 show that Isolation Forest and TCCM share more common features than the Decision Tree on several datasets. For example, on *magic.gamma*, both methods identify Alpha, Width, Length, Conc, and Conc1 among their most important features, although with different ordering. Likewise, on *cover*, both methods highlight Elevation, Slope, and Hillshade. However, the rankings diverge a lot on *celeba* and *UNSW-NB15*, where the two methods emphasize different groups of features.

5.3.3 Discussion

The results suggest that the balance between predictive performance and interpretability depends strongly on the classical model that is used for comparison. While the Decision Tree achieves

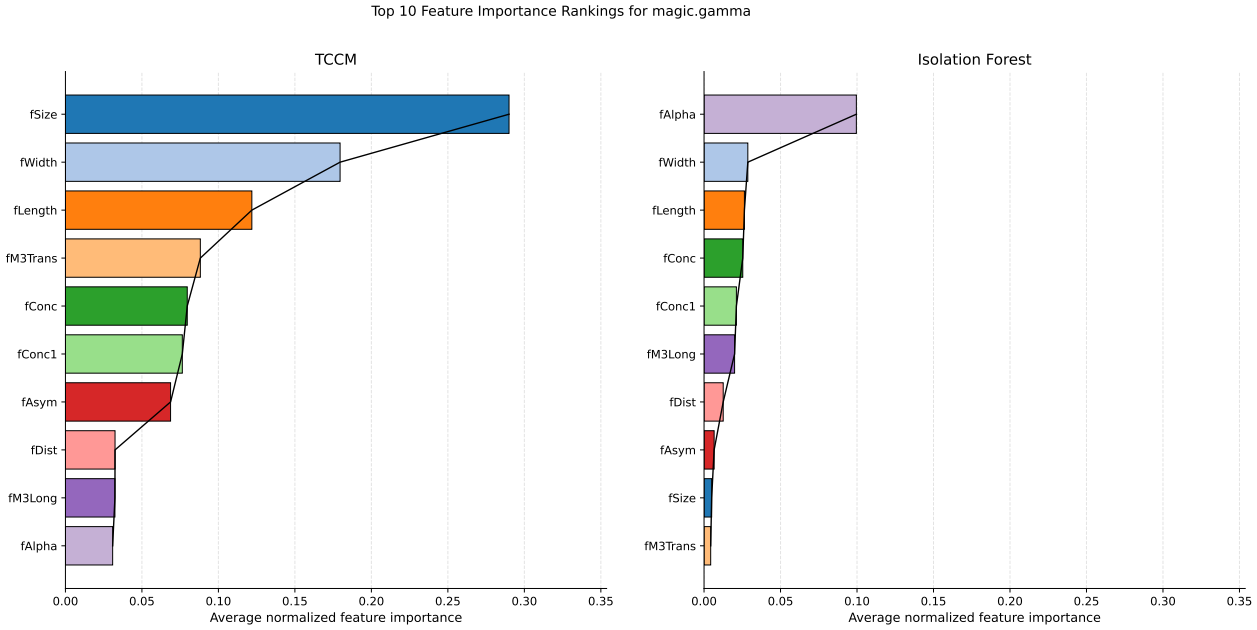


(a) *celeba*

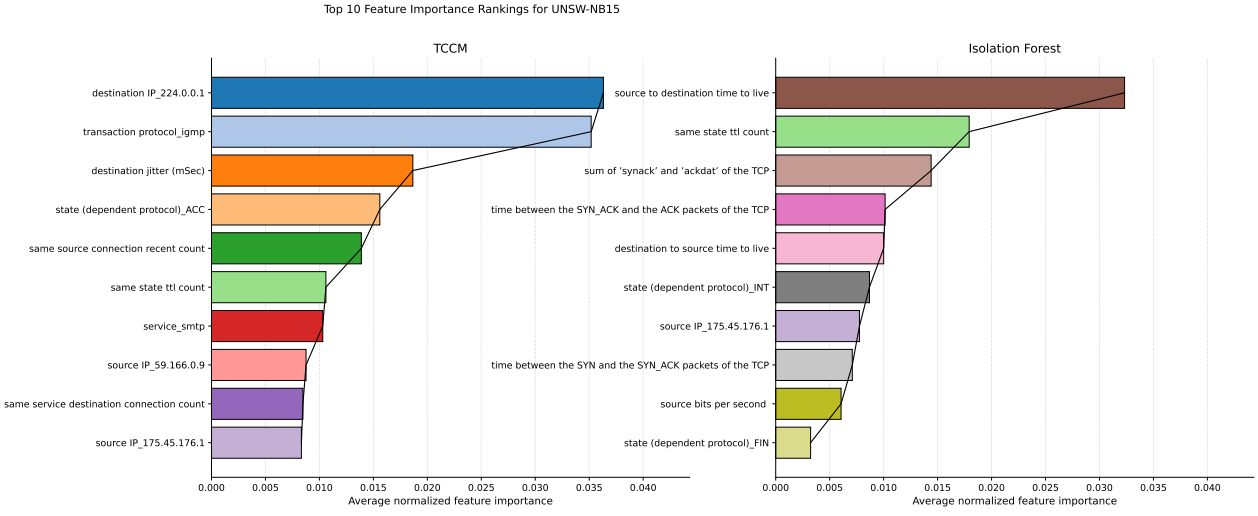


(b) *cover*

Figure 7: Feature ranking comparisons between TCCM and Isolation Forest on the *celeba* and *cover* datasets. The left rankings show TCCM's intrinsic feature importance scores, while the right rankings show Isolation Forest permutation feature importance scores.



(a) *magic.gamma*



(b) *UNSW-NB15*

Figure 8: Feature ranking comparisons between TCCM and Isolation Forest on the *magic.gamma* and *UNSW-NB15* datasets. The left rankings show TCCM's intrinsic feature importance scores, while the right rankings show Isolation Forest permutation feature importance scores.

the highest anomaly detection performance across all datasets, its feature importance rankings generally shows very limited agreement with TCCM's intrinsic explanations. In contrast, Isolation Forest produces lower predictive performance on some datasets but demonstrates slightly stronger agreement with TCCM's feature-wise importance scores.

The Decision Tree consistently outperforms TCCM in terms of AUROC. This indicates that a supervised interpretable classifier can effectively learn different patterns when anomaly labels are available during training (which TCCM did not have access to). However, the explanation rankings reveal that the features used by the Decision Tree often differ significantly from the ones that TCCM emphasizes. The low Jaccard and RBO values across most datasets show that the two approaches rely on different representations of anomalous behaviour. This difference is due to the fact that the Decision Tree optimizes classification accuracy with the help of labelled examples, meanwhile TCCM learns a contraction path of normal data and detects deviations from this learnt structure.

Isolation Forest functions differently. Although its predictive performance is lower than that of the Decision Tree and TCCM, its rankings overlap with TCCM's intrinsic scores more, particularly on the *magic.gamma* and *cover* datasets. This may be explained by the fact that both Isolation Forest and TCCM operate in an anomaly detection setting and identify anomalies as deviations from the structure of normal data rather than with direct supervision.

When interpreting the anomaly detection performance results, it should be noted that the TCCM experiments in RQ3 are not conducted under exactly the same evaluation setup as in RQ1 and RQ2. Consequently, the TCCM AUROC values reported in RQ3 are not directly comparable to those obtained in RQ1 and RQ2. This difference is particularly visible for the *magic.gamma* dataset, where TCCM achieves a lot lower predictive performance in RQ3 than in the earlier experiments. Therefore, part of the observed performance difference may be caused by the modified evaluation setup rather than the changes in the TCCM model itself.

According to RBO scores, the overall agreement remains relatively low across all experiments. The *celeba* dataset again demonstrates weak agreement between explanation methods. Similar to the findings of the first and second research questions, the binary nature of the facial attributes may make the learnt representations more sensitive to differences in model architecture. Consequently, the interpretable models and TCCM focus on different facial attributes when identifying anomalous samples. The *UNSW-NB15* dataset produces very low agreement despite achieving the highest predictive performance for all methods. TCCM frequently highlights protocol, timing, and IP address features, whereas the classical models focused more strongly on connection statistics and network timing features. This may imply that different anomaly detection mechanisms may identify distinct aspects of anomalous network traffic even when they achieve similar classification performance.

When comparing these findings with the first and second research questions, a pattern emerges. The agreement between TCCM and SHAP explanations in RQ1 and RQ2 is higher than the agreement observed between TCCM and the classical interpretable models in RQ3, especially on the higher-dimensional datasets. This is because SHAP explanations are still derived from anomaly scores produced by either TCCM or DTE, while the Decision Tree and Isolation Forest generate both their predictions and explanations independently. Therefore, the rankings in RQ3 reflect differences between separate anomaly detection mechanisms rather than differences between explanation techniques applied to the same underlying model.

To conclude, the results provide only limited evidence that TCCM achieves a better balance

between accuracy and interpretability than classical interpretable machine learning approaches. While the Decision Tree generally achieves higher predictive performance and provides directly interpretable feature importance scores, its explanations show little agreement with TCCM's intrinsic feature-wise scores. Isolation Forest demonstrates stronger agreement with TCCM's explanations on several datasets, although its predictive performance is generally lower. Importantly, some of the observed performance differences may be attributable to the modified experimental setup rather than to the model itself. Overall, no single method dominates both predictive performance and interpretability. TCCM's main advantage remains its ability to provide intrinsic explanations that are directly derived from its internal structure, while the classical interpretable approaches rely on different learning objectives and explanation structures. Consequently, the explanation scores and rankings produced by these methods are not fully comparable, although the observed overlaps indicate that they may still capture some common characteristics of anomalies.

6 Conclusions and Limitations

This thesis set out to investigate the explainability of Time-Conditioned Contraction Matching (TCCM). As deep anomaly detection methods rely on neural networks to achieve strong predictive performance, it has become an important challenge to understand the reasoning that leads to the detected anomalies. TCCM aims to bridge this gap by offering state-of-the-art performance and intrinsic interpretability, which may remove the need for costly post-hoc explanation methods. Therefore, we examined whether TCCM's intrinsic feature-wise explanations are consistent with established post-hoc explanation methods and explored the balance between explainability and anomaly detection performance.

To achieve these goals, three research questions were defined. First, TCCM's intrinsic feature-wise importance scores were compared with SHAP explanations of TCCM anomaly scores across four tabular datasets. Secondly, the intrinsic interpretations of TCCM were compared with SHAP explanations generated for a different deep anomaly detection method, DTE non-parametric. Finally, TCCM was compared with interpretable machine learning models, namely Decision Tree and Isolation Forest, in order to evaluate the trade-off between anomaly detection performance and explainability.

The results of the first research question show that TCCM's intrinsic explanations and SHAP explanations of TCCM predictions vary in terms of agreement depending on the dataset. While moderate agreement was observed on some datasets, large differences were also found in both the selected features and their ranking positions. These findings suggest that TCCM's intrinsic feature importance scores do not necessarily reproduce the feature attributions generated by SHAP, although both methods often highlighted some common patterns. This can have implications for relying on TCCM's own explanations in scenarios when TCCM achieves a high anomaly detection performance. In these cases, TCCM's intrinsic interpretations may provide competing explanations of anomaly scores compared to SHAP explanations of anomalies identified by TCCM. As for which explanations can be considered 'correct', one could argue that TCCM's intrinsic interpretability should enable more truthful explanations as these are derived directly from the model's inner structure. On the contrary, SHAP is more dependent on the background data that defines what a normal sample looks like, and the exact perturbation settings. While determining with a high certainty which explanations are the 'correct' ones requires further, more extensive experiments, these results point to TCCM as a better choice if the goal is to gain interpretations of the inner anomaly scoring mechanism while maintaining computational efficiency and high-performance. On the other hand, these conclusions do not seem to apply to cases where TCCM achieved lower predictive performance. Since the classified anomalies may not actually be anomalies, making the explanation quality questionable, a post-hoc method may prove to be more informative as the explanations are based on how the model output changes under feature perturbation rather than the model's internal anomaly scoring mechanism. For example, future work could explore why TCCM seems to struggle with datasets consisting of exclusively binary features, and the connection between anomaly detection performance and intrinsic explainability.

The second research question demonstrates that the agreement between TCCM's intrinsic explanations and SHAP explanations of DTE predictions was also dependent on the dataset. Here, the results indicate that TCCM and DTE can attribute importance to different features even when they identify similar anomalies. Nevertheless, several datasets show meaningful overlap among the most highly ranked features, which suggests that the two methods may capture related

characteristics of the data despite relying on different internal mechanisms.

The third research question investigates the balance between explainability and predictive performance. The results show that TCCM consistently achieved stronger anomaly detection performance than Isolation Forest, especially on the more complex and high-dimensional datasets. At the same time, TCCM's feature rankings often differed largely from those of Decision Tree and Isolation Forest. Therefore, these findings cannot state with certainty that TCCM achieved a better balance between accuracy and interpretability.

This thesis has some limitations. Firstly, the experiments were conducted on a limited number of datasets, and the evaluation focused on ranking-based explainability metrics. To further explore the differences between TCCM's interpretability and other explainability methods, one could focus on the relation between the feature-wise importance scores, and carry out more statistical tests. Secondly, SHAP explanations were generated for only a subset of anomalous instances in the second research question due to computational constraints. To overcome these limitations, future work could extend the evaluation to additional anomaly detection models, alternative explanation methods, more datasets, and different evaluation metrics. Additionally, if extensive computational resources are available, SHAP could explain more anomalies identified by DTE.

In conclusion, the results indicate that TCCM's intrinsic interpretations partially align with the explanations of established post-hoc explanation methods, but the degree of agreement depends strongly on the dataset and comparison method. Additionally, the comparison with Decision Tree and Isolation Forest showed that higher anomaly detection performance does not necessarily correspond to greater agreement in feature importance rankings. Together, these findings suggest that different explanation approaches may reveal different aspects of a model's decision-making. Therefore, intrinsic explanations should be considered together with established post-hoc explanation techniques, rather than its replacements. While TCCM's intrinsic explanations can provide a guide to understanding the underlying structure of anomaly scoring, it may not capture some explanations that an "outsider" post-hoc explanation method can find. Therefore, the disagreement between the methods should not be interpreted as one of the methods is correct while the other one is incorrect. Instead, TCCM's intrinsic explanations may be recommended as the primary explainer when TCCM is used for anomaly detection and it achieves a high predictive performance, while post-hoc explanation methods may prove to be a better fit for black-box models. Additionally, applying post-hoc models to TCCM when it achieves a lower anomaly detection performance may give a clearer picture of how each feature contributes to the anomalies.

References

[1] Charu C. Aggarwal. *An Introduction to Outlier Analysis*, pages 1–34. Springer International Publishing, Cham, 2017.

[2] Sajid Ali, Tamer Abuhmed, Shaker El-Sappagh, Khan Muhammad, Jose M. Alonso-Moral, Roberto Confalonieri, Riccardo Guidotti, Javier Del Ser, Natalia Díaz-Rodríguez, and Francisco Herrera. Explainable artificial intelligence (xai): What we know and what is left to attain trustworthy artificial intelligence. *Information Fusion*, 99:101805, 2023.

[3] Jock Blackard. Covertype. UCI Machine Learning Repository, 1998. DOI: https://doi.org/10.24432/C50K5N.

[4] R. Bock. MAGIC Gamma Telescope. UCI Machine Learning Repository, 2004. DOI: https://doi.org/10.24432/C52C8B.

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM Comput. Surv.*, 41(3), July 2009.

[6] Tom Fawcett. An introduction to roc analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. ROC Analysis in Pattern Recognition.

[7] Songqiao Han, Xiyang Hu, Hailiang Huang, Mingqi Jiang, and Yue Zhao. Adbench: Anomaly detection benchmark, 2022. arXiv (2206.09426).

[8] Hadi Hojjati, Thi Kieu Khanh Ho, and Narges Armanfard. Self-supervised anomaly detection in computer vision and beyond: A survey and outlook. *Neural Networks*, 172:106106, 2024.

[9] Zhong Li. Tccm-nips. https://github.com/ZhongLIFR/TCCM-NIPS, 2025.

[10] Zhong Li, Qi Huang, Yuxuan Zhu, Lincen Yang, Mohammad Mohammadi Amiri, Niki van Stein, and Matthijs van Leeuwen. Scalable, explainable and provably robust anomaly detection with one-step flow matching, 2025. arXiv (2510.18328).

[11] Fei Tony Liu, Kai Ting, and Zhi-Hua Zhou. Isolation forest. pages 413 – 422, 01 2009. 2008 Eighth IEEE International Conference on Data Mining. DOI: 10.1109/ICDM.2008.17.

[12] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[13] Victor Livernoche, Vineet Jain, Yashar Hezaveh, and Siamak Ravanbakhsh. On diffusion modeling for anomaly detection, 2025. arXiv (2305.18593).

[14] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions, 2017. arXiv (1705.07874).

[15] Christoph Molnar. *Interpretable Machine Learning*. 3 edition, 2025.

[16] Nour Moustafa and Jill Slay. Unsw-nb15: a comprehensive data set for network intrusion detection systems (unsw-nb15 network data set). In *2015 Military Communications and Information Systems Conference (MilCIS)*, pages 1–6, 2015.

[17] Nour Moustafa. Unsw-nb15 dataset. https://research.unsw.edu.au/projects/unsw-nb15-dataset, 2015. Accessed: 2026-05-26.

[18] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM Comput. Surv.*, 54(2), March 2021.

[19] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. "why should i trust you?": Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 1135–1144, New York, NY, USA, 2016. Association for Computing Machinery.

[20] Ribana Roscher, Bastian Bohn, Marco F. Duarte, and Jochen Garcke. Explainable machine learning for scientific insights and discoveries. *IEEE Access*, 8:42200–42216, 2020.

[21] Ahmed M. Salih, Zahra Raisi-Estabragh, Ilaria Boscolo Galazzo, Petia Radeva, Steffen E. Petersen, Karim Lekadir, and Gloria Menegaz. A perspective on explainable artificial intelligence methods: Shap and lime. *Advanced Intelligent Systems*, 7(1):2400304, 2025.

[22] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[23] William Webber, Alistair Moffat, and Justin Zobel. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28:20, 11 2010.