Behavior-Based Interpretable Trust Management for IoT Systems

Muhammad Aaqib
School of computing
Ulster university
Northern Ireland, UK
aaqib-m@ulster.ac.uk

Aftab Ali
School of computing
Ulster university
Northern Ireland, UK
a.ali@ulster.ac.uk

Liming Chen
School of Computer Science and Technology
Dalian University of Technology
Dalian, China
limingchen0922@dult.edu.cn

Omar Nibouche
School of computing
Ulster university
Northern Ireland, UK
o.nibouche@ulster.ac.uk

***Abstract*—** Establishing appropriate trust at the "right" level for the "right" application at the "right" time is important in the constantly evolving environment of the Internet of Things (IoT). Nevertheless, the process is challenging due to the lack of explainability and interpretability of the machine learning models. This paper presents a novel approach to managing IoT trust by employing explainable artificial intelligence (XAI) to connect complex algorithmic decisions with human understanding. Specifically, we propose a mutual information selection technique to determine the most significant behaviour-based features to identify trustworthy and untrustworthy behaviour in IoT device systems. Based on these behaviour-based features we develop a rule-based decision tree (DT) method to help enhance the explainability of our model. We evaluate our approach with a transformed UNSW NB 15 dataset, the results demonstrate improved user trust and system transparency. In addition, we did a comparative analysis of our BB-TMS with state-of-the-art methods, which demonstrated that our model surpasses competitors in terms of precision and interpretability. This highlights the effectiveness of integrating XAI with traditional machine learning approaches in the IoT domain.

***Keywords*—** *BB-TMS, IoT, XAI, Rule-based, DT*

I. INTRODUCTION

The Internet of Things (IoT) has brought about a substantial transformation in connectivity, facilitating the interconnection of diverse devices across several domains. [1] [2]. With the integration of IoT into our daily lives, it is crucial to prioritise ensuring the security of these interconnected systems from malicious behaviours. An essential aspect of protecting these systems is the establishment of efficient trust management system (TMS) in IoT device systems, which guarantees the dependability and authenticity of these extensive networks. Traditional TMS, although somewhat effective, frequently lacks the required transparency [3] [4] and interpretability [5] that are essential for comprehending and improving security measures.

The objective of our research is to address a notable gap by introducing a novel behavioral based (BB) TMS for IoT device systems. This approach is based on the fundamentals of explainable artificial intelligence (XAI). Moreover, the main part of our model is a rule-based decision tree (DT) classifier, selected for its outstanding knowledge of interpretability and effectiveness. Our BB-TMS demonstrates the XAI concept by providing clear and comprehensible decision-making methods, which are essential for establishing trust among IoT device systems. In this paper, we focus on selecting BB features through a mutual information (MI) technique. This is critical for determining the most significant BB features to identify trustworthy and untrustworthy behavior in IoT device systems. By enhancing the accuracy of our BB-TMS, we are able to follow the objectives of XAI, which aim to enhance transparency in AI-based decision-making. Expanding upon this basis, we conduct a comparison between our rule-based DT, Support Vector Machine (SVM), and Naive Bayes (NB) models. This comparison demonstrates the BB-TMS's interpretability, reliability, and superior performance in trust establishment, highlighting its effectiveness in enhancing IoT security. This study's notable contributions include the following:

- We present a novel BB-TMS for IoT systems, employing rule-based DT XAI to improve the transparency and interpretability of TMS. This system represents a notable change from traditional methods, providing a more advanced and comprehensible approach for establishing trust.
- This study employed MI to choose BB features and determine their level of impact on trust establishment**.**
- A rule based explanation makes our BB-TMS more transparent and improves user trust.

The remaining sections of the paper are organised as follows: Section II presents an overview of the traditional TMS in IoT device systems. Section III presents a comprehensive overview of the proposed BB-TMS. Subsequently, in Section IV, we provide and analyse the results and the comparative analysis that has been produced and implemented. Finally, the paper is briefly summarised in Section V.

II. REALTED WORK

The literature review aims to outline the research conducted on TMS in IoT device systems. Previous research has shown that researchers have made significant efforts to develop TMS using ML approaches. The investigation's findings are summarised as follows:

Wang et al. [6] present ML-based TMS in Internet of Vehicles. The developed system distinguishes between trustworthy and non-trustworthy vehicles by utilising an appropriate threshold. The study conducted by Marche et al. [7] focuses on the development of feedback mechanisms and

the implementation of different evaluation metrics. In addition, they introduced a new collusive attack that impacts the evaluation of the services received. By conducting simulations with a real IoT dataset, the authors showed the importance of producing a response and the impact of the recently introduced malicious behaviour. Another study [8] presented a novel model for identifying trust-related threats in IoT devices. The model leverages deep Long Short-Term Memory (LSTM) model. The purpose of the concept was to identify and separate untrustworthy behaviours within IoT systems. The model attained accuracy and F-measure scores of 0.98 and 0.99, respectively. Moreover, the authors of [9] developed a TMS using ML models to distinguish between trustworthy and non-trustworthy devices. The authors used a transformed UNSW-NB15 and claimed accuracy and precision of 0.92 and 0.983 respectively. Furthermore, the authors of [10] investigated the XAI- for TMS and employed a DT-based random forest (RF) to tackle the problem. Through the implementation of Explainable Artificial Intelligence (XAI) on these intricate AI models, the authors claim an accuracy and F1_score of 0.98.

The literature study indicates that a significant number of investigations have been carried out in TMS [6] [7] [8] [9], exploring various methods of ML for prediction of malicious behaviour in IoT device systems. A crucial problem in these systems is that the authors couldn't provide any valuable understanding of the explainability of the models employed in TMs. This gap highlights a crucial part of trust management in IoT networks, comprehending not only the mechanisms through which these models arrive at their decisions but also the underlying reasons behind them. It is crucial to be able to understand and provide valid reasons for the decisions made by TMS in order to provide transparency, dependability, and user confidence. Without this degree of comprehensibility, implementing these models in practical situations may face uncertainty, impeding their acceptability and efficacy. Therefore, we have developed a BB-TMS that enhances the model's explainability using rule-based analysis. Moreover, we precisely developed our system to efficiently evaluate the trustworthiness of IoT device systems and offer clear and comprehensible explanations for their evaluations. This ensures that users can easily understand the reasoning behind its trust evaluations.

III. PROPOSED METHODOLOGY

This section provides an overview of the ML techniques for classifying the trustworthy and non-trustworthy classes. This includes an in-depth review of the dataset used, the preprocessing of the data, and the model prediction using Explainable Artificial Intelligence (XAI). In addition, we proposed the XAI-BB-TMS for IoT device systems. The approach involves the mutual information (MI) based FS and integration of ranking based feature importance using DT with the interpretation of the proposed model based on rules, as depicted in Fig. 1. The main contribution of the developed model is to identify and classify the network traffic and label them as trustworthy and untrustworthy class. The process of classification can be divided into two distinct steps. Stage 1 encompasses various tasks, namely data preprocessing, features engineering and finally employing the three ML models, specifically DT, SVM and NB. Moreover, in stage 2, the importance of ranking-based features is determined using a DT model. Subsequently, the model is interpreted by employing a rules-based interpretation approach.

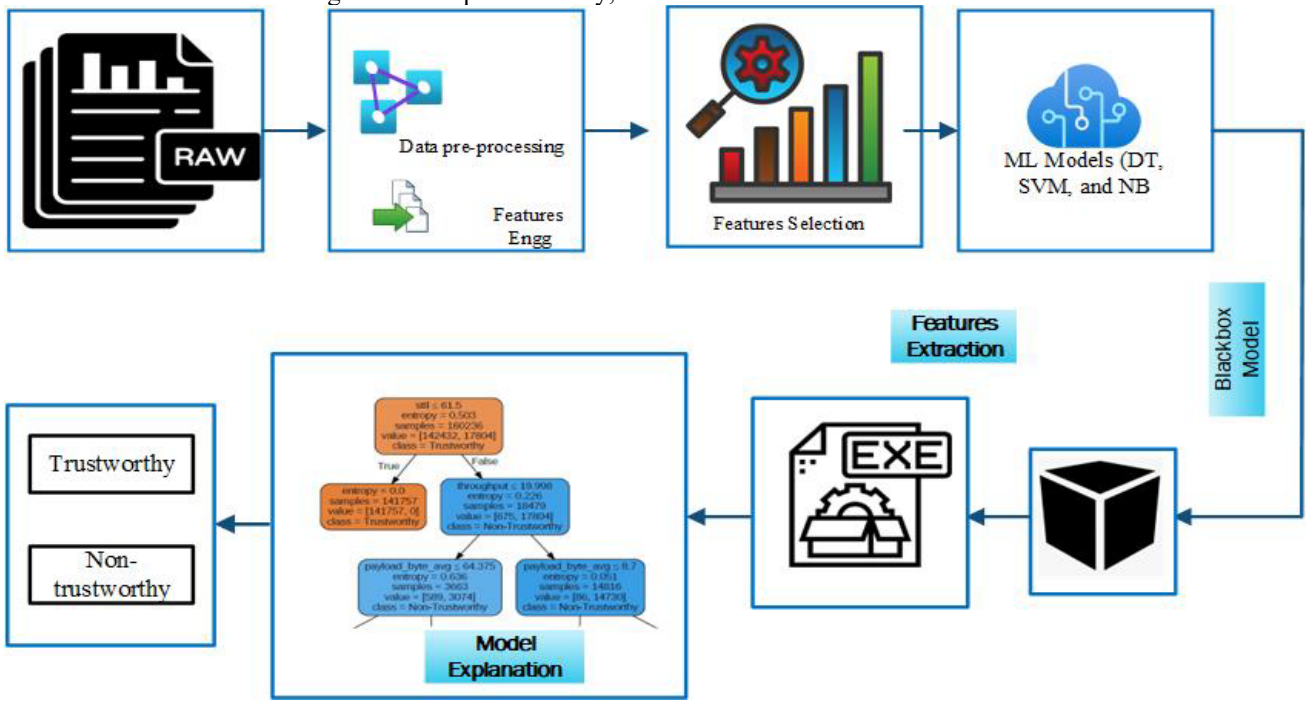


Fig 1. Proposed framework for explainable artificial intelligence trust management system (XAI-TMS)

A. Dataset and Preprocessing

In this paper we choose raw and unprocessed data of UNSW-NB15 dataset as the foundation of our analysis and to achieve the maximum level of trustworthiness. In addition,

we went through an extensive feature engineering process to enhance the dataset's inherent potential. This rigorous feature engineering effort resulted in the development of numerous BB features. These newly obtained BB features were critical in expanding our dataset and, as a result, facilitating the extraction of useful insights. The label encoding approach provides each BB feature with a distinctive integer value. The subsequent procedure involves replacing the empty, dash, and infinite attribute values with 0 to obtain a dataset consisting solely of numerical values. The label-encoding approach transforms the category values in an attribute into an index. For instance, the label-encoding approach converts the categorical value of an ICMP protocol into a TCP value. This approach enhances the efficiency of the model. Moreover, the model transforms classes like trustworthy and untrustworthy into binary digits, substituting them with 0 and 1, respectively.

B. Features Selection

Mutual information (MI) is a measure that quantifies the level of dependency or shared knowledge between two random variables. Unlike the correlation coefficient, which can only evaluate linear dependencies, MI has the ability to detect both linear and nonlinear correlations between variables. This attribute has played a significant role in its extensive acceptance for feature selection in the references [9] [11] [12] [13] [14] [15]. Due to the complex and multifaceted features of the transformed UNSW NB 15 dataset, it is crucial to identify the most important features in order to improve the performance of the model and ensure its interpretability. In order to accomplish this, we employ a feature selection technique based on Mutual Information (MI). This method calculates the level of interdependence between variables, specifically assessing the amount of information that can be obtained about one random variable by observing another. Within our analysis, MI is utilised to ascertain the correlation between each feature and the target variable. This allows us to find the aspects that have a significant impact in detecting network behaviours. This strategy efficiently reduces the complexity of the dataset, enabling the model to focus on qualities that have the greatest predictive ability. In section 4, we will examine the process of ranking the characteristics based on their value, as decided by information gain (IG). This will help us to further enhance the focus and effectiveness of our BB-TMS.

C. Model Deployment

The decision tree method is a notable nonparametric approach within the domain of supervised learning approaches. The use of a tree-like structure for decision-making helps to clarify the possible different outcomes of instances. This approach produces decision rules that possess a high level of comprehensibility and integrate if-then statements. Unlike other supervised models like SVM, and NB, DT offers the advantage of facilitating visual representation and knowledge of the underlying data logic. Moreover, the DT tree employs splitting criteria in order to generate rules. Multiple strategies exist for creating decision trees, and among them, but here we employ the ID3 algorithm, which stands for Iterative Dichotomiser. The ID3 algorithm does a top-down, heuristic search on the given training datasets in order to assess each feature at every node. Furthermore, we incorporate a statistical metric known as IG to determine the optimal BB features to evaluate at each node in the tree-like structure. This statistical measure evaluates the effectiveness of a BB feature in accurately categorising training cases based on their desired classifications. The analytical process is dependent on the concept of entropy.

Entropy is a metric derived from the concept of IG that serves to quantify the level of impurity present within a given dataset (D). Mathematically;

$$E(D) = \sum_{i=1}^{k} p_i log_2(p_i) \qquad 1$$

Where E shows the measure of entropy, while pi represents the probability of the D being classified into class i. The log with a base of *2* indicates that entropy measures the amount of information needed to encode data, expressed in bits.

Information Gain (IG) It is employed to quantify the expected reduction in entropy when a set of samples is divided based on a certain set of features. Mathematically;

$$IG(D,S) = E(D) - \sum_{v \in S} \frac{|D_v|}{|D|} E(D_v) \qquad 2$$

In this particular situation, the collection of all potential values for feature S is referred to as values S , and D_v denotes the subset of D where feature attribute S has the value v. This metric is used to prioritise features and build our DT model.

IV. EXPERIMENTAL ANALYSIS

The suggested model utilises a rule-based decision-making process, incorporating BB features to improve the precision and interpretability of the TMS in IoT environments. The development of this model was conducted using a high-performance PC equipped with an Intel Core™ i9-9900X CPU working at a frequency of 3.50GHz, and 32 GB of RAM.

A. BB Features Ranking based on IG

After utilising the mutual information approach to choose the BB features, the DT algorithm evaluates the training dataset to determine the most efficient feature for distinguishing between trustworthy and untrustworthy instances. This evaluation is based on entropy measurements. Furthermore, the assessment of trustworthiness in a TMS for IoT devices relies on the ranking of several BB features based on their IG. The BB features such as 'throughput', 'sttl', and 'network_latency' are placed at the highest level, indicating their significant impact on the efficiency and trustworthiness of IoT devices. High throughput demonstrates the capacity of devices to process a significant volume of data, whereas the metrics 'sttl' and 'network_latency' can indicate the effectiveness and timeliness of communication, which can influence trust rankings. Conversely, features that offer a smaller amount of information, such as particular payload bytes, acquire a lesser amount of information. This suggests that they have a minor impact on trust-related decisions. The payload bytes may lack a strong association with security or normal activities, indicating that they have a relatively minor impact on the evaluation of trustworthiness. By giving priority to traits that have higher IG, it becomes feasible to calculate trust scores with more precision. Consequently, this aids in establishing a more robust and reliable IoT system by

detecting device behaviour. Fig.2 illustrates the resulting hierarchy of significant BB features. The low significance value of a feature does not necessarily imply its lack of importance for prediction; rather, it suggests that the feature was not chosen early in the tree's development. The trait may be either identical to or strongly linked with another informative feature. BB feature significance levels do not indicate the specific class they are highly predictive for or the relationships between features that may influence prediction.

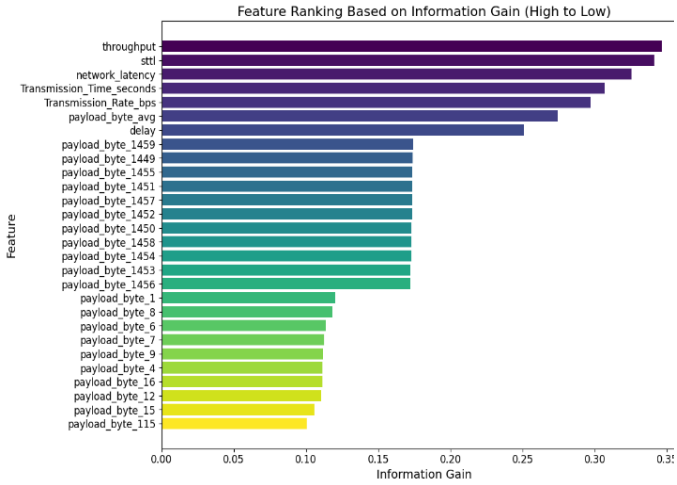


Fig.2 BB features ranking based on IG

B. Improving interpretability using rule-based DT's

The interpretability of DT classifiers is crucial in clarifying the decisions made by ML models, particularly when used in the complex environment of IoT. The investigation will be focused on identifying the reasons behind categorising particular nodes as trustworthy or untrustworthy, using DT rules established from the BB features as depicted in Fig. 3. Moreover, nodes in the DT that have low entropy values are considered trustworthy, indicating a consistent pattern in the data that corresponds to the expected behaviour of IoT devices. For instance, when a node threshold like "payload_byte_avg ≤ 13.07" is used to classify trustworthiness and has an entropy of 0.121, it indicates that specific thresholds can be dependable indications of trustworthy communication. Another example of node_12 is considered trustworthy because it meets the requirements of having a delay ≤ 5.2 while having an entropy of 0.12. The DT recognises this condition as a reliable indicator of trustworthiness, likely due to the fact that instances with little delay are generally trustworthy. Furthermore, node_2 with an entropy of 0.22 is categorised as non-trustworthy according to the criterion that the value of throughput is less than or equal to 19.9. The DT classifies instances with low throughput as untrustworthy, presumably due to network congestion or poor connectivity, which could result in untrustworthy behavior. Also, node_3 has a reasonably high impurity, as shown by its entropy value of 0.636. If the condition payload_byte_12 is less than or equal to 227, the instances are classified as untrustworthy. The node's impurity indicates some misclassification, despite identifying cases where specific payload byte features are linked to untrustworthy behaviour.

BB-TMS in IoT device frameworks, using rule-based DT classifiers, improves our understanding and model interpretability. It also presents an advanced approach to distinguishing between trustworthy and untrustworthy device behaviours based on BB features. By distinguishing nodes with low entropy that are trustworthy from those with higher entropy that are not trustworthy and categorising instances as either trustworthy or untrustworthy based on BB features, these classifiers emphasise the significance of TMS in upholding the trustworthiness of the IoT environment. Examining the classifiers' decision-making process in depth improves the understandability and interpretability of ML applications in the IoT, highlighting the importance of our proposed BB-TMS.

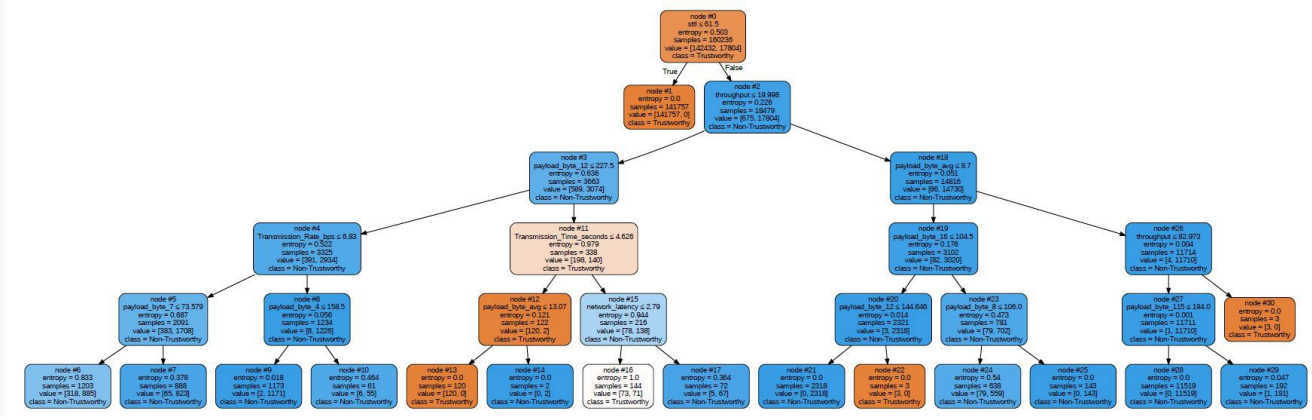
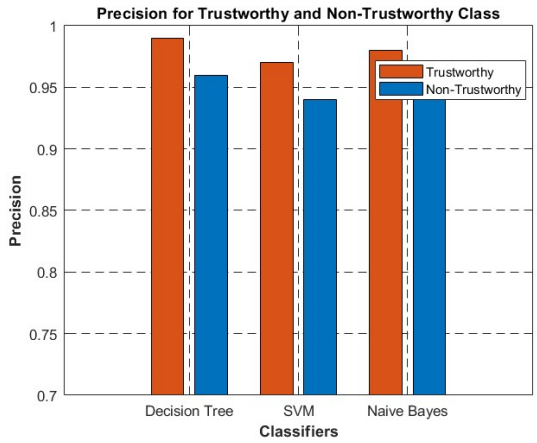


Fig 3. Rule-based outcome from training (Blue nodes are untrustworthy and orange nodes are trustworthy

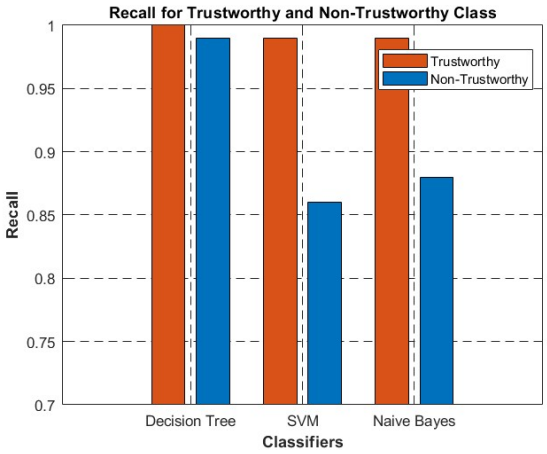
C. Performance Evaluation

To establish trust and detect trustworthy and untrustworthy behavior or devices, we examined the model's effectiveness based on precision, recall, F1-score, and Matthews Correlation Coefficient (MCC) score, as depicted in Fig. 4 (a-c) and Fig.5 . During our investigation, we carefully assessed the performance of the proposed systems and compared the results for each model, such as SVM and NB. After comparative analysis, we found that the rule-based DT model had higher precision and recall scores compared to the NB and SVM models, as depicted in Fig. 4 (a) and (b).



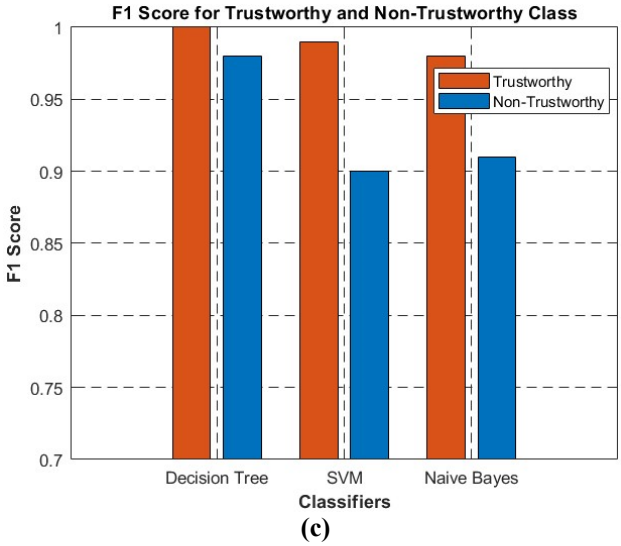
(a)

Furthermore, the F1 scores showed that the rule-based DT model outperformed the others in terms of detecting non-trustworthy behavior or devices, as depicted in figures 4 (c).



(b)

This capability highlights rule-based DT excellence in TMS, as the metrics directly influence the system's efficacy and decision-making process.



(c)

Fig. 4 Performance evaluation of trustworthy and untrustworthy based on (a) Precision (b) Recall and (c) F1_Score

Fig.5 illustrates the MCC scores of three distinct ML models employed in TMS. The DT model demonstrates a remarkably high MCC score, slightly below 1, which indicates its strong ability to detect untrustworthy behaviours. The NB model closely follows, achieving an MCC of 0.9, indicating excellent performance in accurately identifying untrustworthy behaviors in IoT device systems. Furthermore, the SVM model demonstrates a lower score of 0.88 than the other two models. Therefore, these findings indicate that the rule-based DT model may be more effective in establishing trust and identifying untrustworthy behaviour in an IoT device systems.

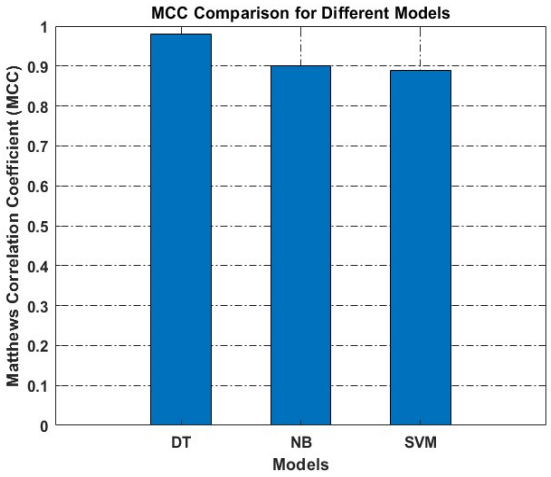


Fig. 5 Models evaluation based on Matthews Correlation Coefficient

D. Comparative Analysis with State of the Art Methods

In Table 1, we compared the performance metrics of various models for an XAI-TMS. These comparative analyses reveal significant differences in performance when evaluated using recall, precision, F1 score, and MCC. These differences directly affect establishing trust and identifying untrustworthy devices. The scheme developed by Iftikhar et al. [16] robustly identifies non-malicious IoT devices and effectively reduces false positives, as indicated by its high recall and precision values, approaching 0.98. Nevertheless, the absence of F1 scores and MCC metrics impedes a thorough understanding of its effectiveness in detecting malicious behaviour. Mane et al. [17] model demonstrates a precision of 0.96, indicating a high level of precision. However, it has a lower recall of 0.712, suggesting that it may not recognise all trustworthy behaviour. The F1 score of 0.822 provides an understanding of the model's balance between precision and recall. In contrast, Novo et al. [18] approach demonstrates outstanding recall and precision, achieving scores of 0.99 and 0.98, respectively. Likewise, a study conducted by Islam et al. [19] demonstrates robust performance, indicating a dependable approach for ensuring the trustworthiness of the system.

Finally, our XAI-TMS surpasses previous models in terms of recall and precision, achieving an exceptional score of 0.99. Our model also demonstrates an F1 score of 0.98 and an MCC of 0.98. The results emphasise the exceptional trustworthiness and effectiveness of the

proposed XAI-TMS in identifying and detecting untrustworthy behavior of IoT device systems.

TABLE 1. Comparative analysis with state-of-the-art schemes

Authors	Recall	F1 score	Precision	MCC
Iftikhar et al. [16]	0.98	NA	0.98	NA
Mane et al. [17]	0.71	0.82	0.96	NA
Novo et al. [18]	0.99	0.98	0.98	NA
Islam et al. [19]	0.94	0.95	0.95	NA
XAI-BB-TMS	0.99	0.98	0.99	0.98

V. CONCLUSION AND FUTURE WORK

Currently, TMS plays an important role in detecting untrustworthy behaviour across various IoT systems, emphasising the need for models that can accurately distinguish and detect such behaviours. To address this problem, this study proposes a BB-TMS that employs ML techniques, including DT, SVM, and NB. Additionally, we employ the mutual information method for BB feature selection and utilise an information-gain technique by ranking the features based on their informational significance, thereby enhancing the model's performance and comprehensibility. The paper highlights the significance of trust in improving the effectiveness of human-machine interactions, with a notable precision and MCC score of 0.99 and 0.98. It also highlights the crucial role of the rule-based model in clarifying model predictions, which helps in establishing appropriate systems. Moreover, this work represents notable progress in the field of IoT device systems by combining rule-based techniques with a focus on transparency and interpretability in the decision-making process. This demonstrates the usefulness of combining XAI and traditional ML algorithms in IoT device systems. In the future, it will be more interesting to explore advanced interpretability models such as local interpretable model-agnostic explanations and shapely additive explanations and counterfactual explanations. These model-agnostic techniques can provide more nuanced and precise understandings into the decision-making process of XAI-TMS, enabling security analysts to identify the exact features contributing to a decision.

References

[1] H. Allioui and Y. Mourdi, "Exploring the full potentials of IoT for better financial growth and stability: A comprehensive survey," *Sensors,* vol. 23, *(19),* pp. 8015, 2023.

[2] O. Vermesan, P. Friess, P. Guillemin, M. Serrano, M. Bouraoui, L. P. Freire, T. Kallstenius, K. Lam, M. Eisenhauer and K. Moessner, "IoT digital value chain connecting research, innovation and deployment," in *Digitising the Industry Internet of Things Connecting the Physical, Digital and VirtualWorlds*Anonymous River Publishers, 2022, pp. 15-128.

[3] S. R. Sindiramutty, C. E. Tan, S. P. Lau, R. Thangaveloo, A. H. Gharib, A. R. Manchuri, N. A. Khan, W. J. Tee and L. Muniandy, "Explainable AI for cybersecurity," in *Advances in Explainable AI Applications for Smart Cities* Anonymous IGI Global, 2024, pp. 31-97.

[4] S. R. Konda, "Ensuring Trust and Security in AI: Challenges and Solutions for Safe Integration," *International Journal of Computer Science and Technology,* vol. 3, *(2),* pp. 71-86, 2019.

[5] V. Hassija, V. Chamola, A. Mahapatra, A. Singal, D. Goel, K. Huang, S. Scardapane, I. Spinelli, M. Mahmud and A. Hussain, "Interpreting black-box models: a review on explainable artificial intelligence," *Cognitive Computation,* vol. 16, *(1),* pp. 45-74, 2024.

[6] Y. Wang, A. Mahmood, M. F. M. Sabri, H. Zen and L. C. Kho, "MESMERIC: Machine Learning-Based Trust Management Mechanism for the Internet of Vehicles," *Sensors,* vol. 24, *(3),* pp. 863, 2024.

[7] C. Marche, L. Serreli and M. Nitti, "Analysis of feedback evaluation for trust management models in the Internet of Things," *IoT,* vol. 2, *(3),* pp. 498-509, 2021.

[8] Y. Alghofaili and M. A. Rassam, "A Dynamic Trust-Related Attack Detection Model for IoT Devices and Services Based on the Deep Long Short-Term Memory Technique," *Sensors,* vol. 23, *(8),* pp. 3814, 2023.

[9] M. Aaqib, A. Ali, L. Chen and O. Nibouche, "Discriminative features-based trustworthiness prediction in IoT devices using machine learning models," in *2023 IEEE Smart World Congress (SWC),* 2023, pp. 1-6.

[10] H. Mankodiya, M. S. Obaidat, R. Gupta and S. Tanwar, "XAI-AV: Explainable artificial intelligence for trust management in autonomous vehicles," in *2021 International Conference on Communications, Computing, Cybersecurity, and Informatics (CCCI),* 2021, pp. 1-5.

[11] H. Zhou, X. Wang and R. Zhu, "Feature selection based on mutual information with correlation coefficient," *Appl. Intell.,* vol. 52, *(5),* pp. 5457-5474, 2022.

[12] F. Macedo, R. Valadas, E. Carrasquinha, M. R. Oliveira and A. Pacheco, "Feature selection using Decomposed Mutual Information Maximization," *Neurocomputing,* vol. 513, pp. 215-232, 2022.

[13] I. C. Covert, W. Qiu, M. Lu, N. Y. Kim, N. J. White and S. Lee, "Learning to maximize mutual information for dynamic feature selection," in *International Conference on Machine Learning,* 2023, pp. 6424-6447.

[14] S. Liu and M. Motani, "Improving Mutual Information based Feature Selection by Boosting Unique Relevance," *arXiv Preprint arXiv:2212.06143,* 2022.

[15] S. Lall, D. Sinha, A. Ghosh, D. Sengupta and S. Bandyopadhyay, "Stable feature selection using copula based mutual information," *Pattern Recognit,* vol. 112, pp. 107697, 2021.

[16] I. Ahmad, M. Basheri, M. J. Iqbal and A. Rahim, "Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection," *IEEE Access,* vol. 6, pp. 33789-33795, 2018.

[17] S. Mane and D. Rao, "Explaining network intrusion detection system using explainable AI framework," *arXiv Preprint arXiv:2103.07110,* 2021.

[18] X. Larriva-Novo, C. Sánchez-Zas, V. A. Villagrá, A. Marín-Lopez and J. Berrocal, "Leveraging Explainable Artificial Intelligence in Real-Time Cyberattack Identification: Intrusion Detection System Approach," *Applied Sciences,* vol. 13, *(15),* pp. 8587, 2023.

[19] M. T. Islam, M. K. Syfullah, M. G. Rashed and D. Das, "Bridging the Gap: Advancing the Transparency and Trustworthiness of Network Intrusion Detection with Explainable AI," 2023.