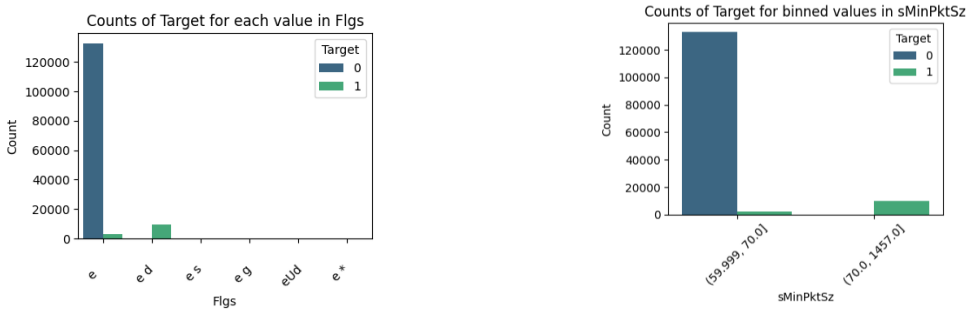
SHIELD: Smart Healthcare Intrusion Detection Using SafeML for IoMT Security

1 Methodology

The proposed methodology, termed SHIELD, presents a comprehensive and safety-aware ML pipeline tailored for intrusion detection using two real-world datasets. It begins with meticulous data preprocessing, involving the elimination of low-variance and biased non-numeric features, label encoding of categorical attributes, and outlier removal using Isolation Forest. Skewed numeric features are normalised using a quantile transformer to improve data distribution. These preprocessing techniques were chosen based on multiple tests using different techniques. Following this, SHIELD employs an AGRM novel auto-weighted feature selection technique that minimises feature redundancy while preserving relevance to select the most informative features. These features are standardised and split for model training and evaluation. Before training, the data undergoes rigorous safety checks to ensure compliance with SafeML standards. A Histogram-Based Gradient Boosting Classifier is then fine-tuned using Grid Search and evaluated. Notably, SHIELD's modular design allows it to integrate seamlessly with various ML strategies and supports human-in-the-loop validation, reinforcing model safety and adaptability across domains.

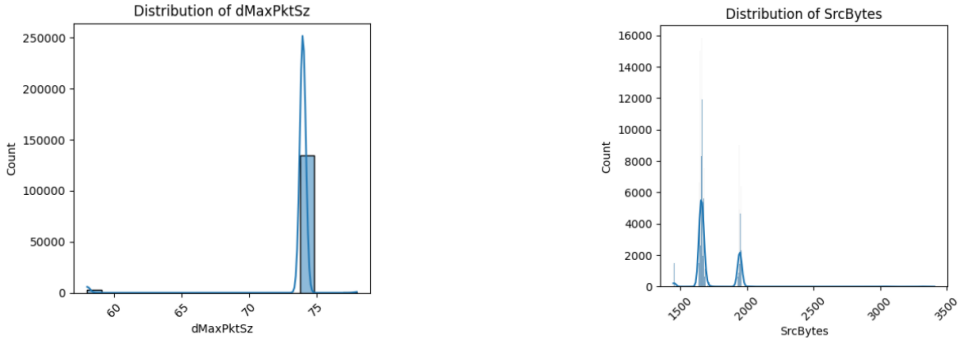
Fig. 1 describes the examples of some of the relationships of some dropped columns with respect to the target variables for D1 dataset. It could be seen that upon inclusion of these, the model would have become biased, say Flgs for example, the model could have categorised all the 'e d' to class 1 without considering the weightage of other features. Finally, these non-numeric features are label encoded to convert them into the numeric format.

For the numeric features, an isolation forest was applied for outlier detection. This method helped remove 7257 outliers from the D1 data and 816 from the D2 dataset. Finally, a feature distribution for each feature was plotted to understand the data distribution.This helped to identify the skewed or tailed features such as `SrcBytes`, `SrcGap`, `SrcLoad`, `DstLoad`, `SIntPkt`, `SIntPktAct`, `DIntPkt`, `SrcJitter`, `DstJitter`, `Dur`, `dMaxPktSz`, `dMinPktSz`, `Loss`, `pLoss`, `pSrcLoss`, `Rate`, `svmem_percent`, and `network_load` from dataset D1, and `SrcLoad`, `DstLoad`, `SIntPkt`, `DIntPkt`, `SrcJitter`, `DstJitter`, and `Dur` from dataset D2. Fig. 2 describes an example distribution of these columns. The quantile transformer was used to remove the skewness of these features. Fig. 3 describes the distribution of these columns after passing through the transformer.



(a) Flgs vs target (b) sMinPktSz vs target

Fig. 1: columns vs target for WUSTL-EHMS-2020 dataset



(a) Distribution of dMaxPktSz (b) Distribution of SrcBytes

Fig. 2: Feature Distribution before quartile transformation for WUSTL-EHMS-2020 dataset

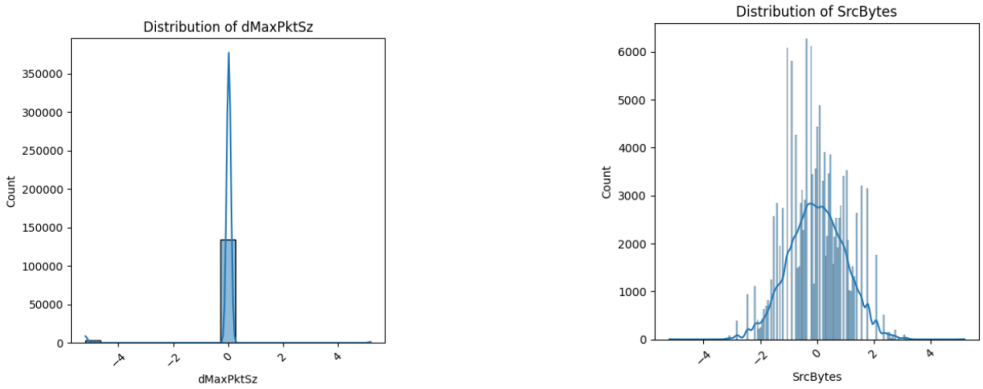
2 Experimental Evaluation

2.1 Results analysis

This section describes the results obtained on both the D1 and D2 datasets. For simplicity, it initially discusses the results obtained through the binary classification and then describes the multi-class classification.

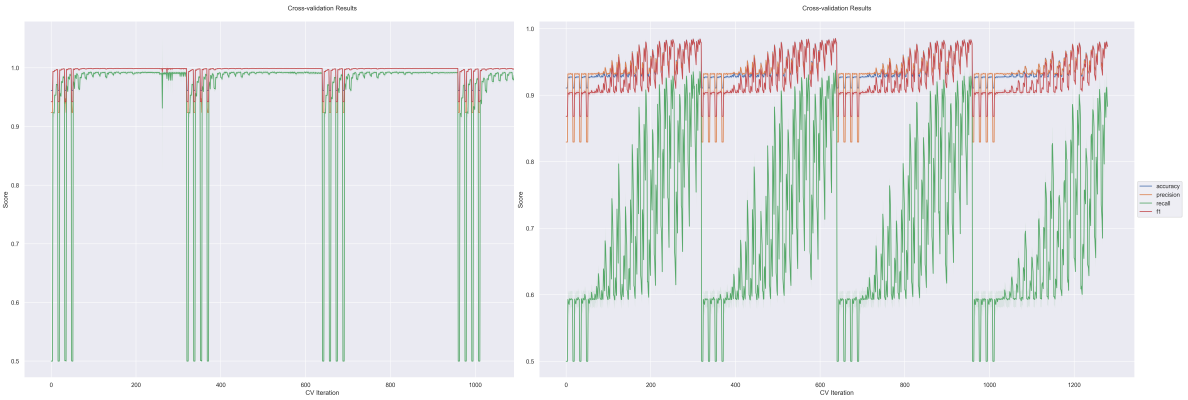
2.1.1 Binary classification

Notably, while models like DT and KNN demonstrate high performance with accuracies around 97.55% and 94.58%, respectively, the Perceptron model lags behind with an accuracy of only 69.52%, even though its precision is relatively high at 84.42%. This shows that besides implementing the safe techniques into the proposed method, SHIELD performs better than the other ML and DL models.



(a) Distribution of dMaxPktSz (b) Distribution of SrcBytes

Fig. 3: Feature Distribution after quartile transformation for WUSTL-EHMS-2020 dataset

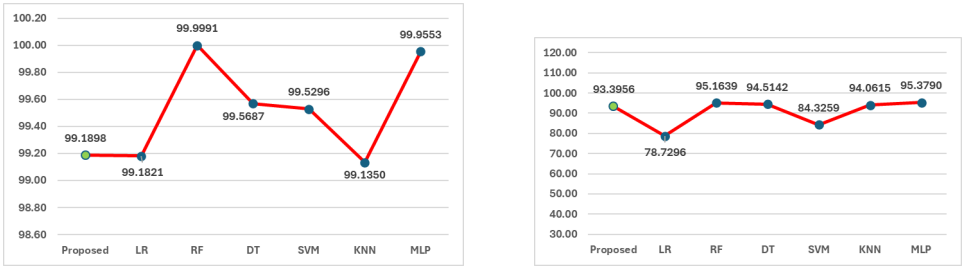


(a) WUSTL-HDRL-2024 dataset (b) WUSTL-EHMS-2020 dataset

Fig. 4: CV across two datasets for binary classification

SHIELD records an AUC of 99.19%, which is very competitive with the traditional models, with LR scoring nearly identical at 99.18% as seen in Fig. 5. Notably, the RF model achieves an almost perfect performance with an AUC of 99.9991%, while the MLP also performs remarkably high at 99.9553% as seen in Fig. 5a. But Fig. 6 indicates that SHIELD takes lesser time in comparison to the models outperforming it. For instance, RF is 30 times slower as compared to SHIELD and MLP is around 80 time slower as compared to SHIELD. Moreover, KNN also takes 10 time more time than the SHIELD, highlighting the supremacy of the SHIELD architecture.

Fig. 5b, SHIELD attains an AUC of 93.40%, substantially outperforming LR (78.73%) and SVM (84.33%). RF again leads with the highest score at 95.16%, and



(a) WUSTL-HDRL-2024 dataset (b) WUSTL-EHMS-2020 dataset

Fig. 5: ROC values across two datasets for binary classification

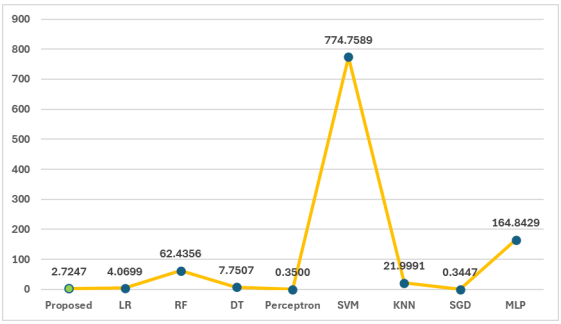


Fig. 6: Average time comparison of various model on WUSTL-HDRL-2024 dataset for binary classification

both DT and KNN yield AUCs in the mid-90s. The MLP, with an AUC of 95.38%, demonstrates strong discriminative power comparable to RF. As seen with D1, likewise Fig. 7 depicts the average training time for these models, indicating that RF, DT, and MLP take more time in comparison to the SHIELD.

Fig. 8 presents the Mean Absolute Error (MAE) values for various models evaluated on a binary classification task. SHIELD achieved an MAE of 0.0009, demonstrating competitive performance in comparison to standard models. Notably, the RF model yielded the lowest MAE of 0.0006 as seen in Fig. 8a, but it needs more computation time. Fig. 8b indicates that SHIELD achieved the lowest MAE of 0.0135, outperforming all baseline models. This performance demonstrates robustness and adaptability. Traditional models reported higher error rates of 0.0745, 0.0726, and 0.0748, respectively, indicating lower prediction accuracy in this context. Ensemble and neural-based methods also showed improved results compared to linear models, with MAEs of 0.0591, 0.0518, and 0.0542, respectively. The Perceptron model performed the worst, with an MAE of 0.3048, suggesting poor generalisation on this dataset. Overall, the results confirm the effectiveness of the SHIELD, especially when applied to diverse data distributions in binary class scenarios.

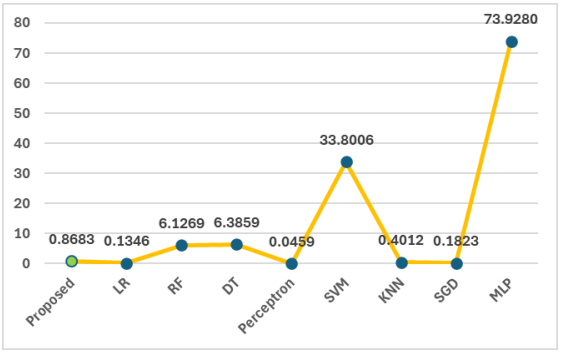
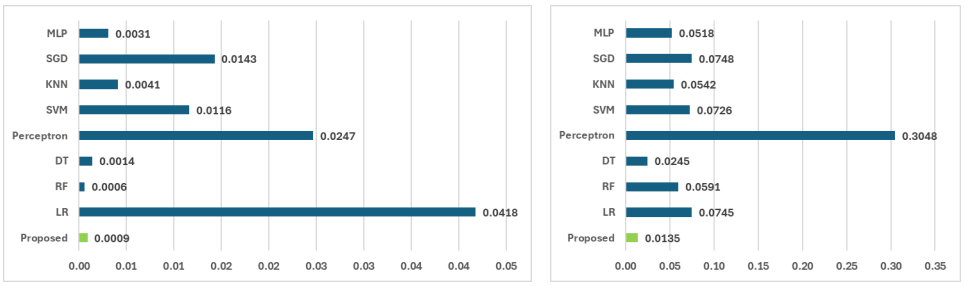


Fig. 7: Average time comparison of various model on WUSTL-EHMS-2020 dataset for binary classification

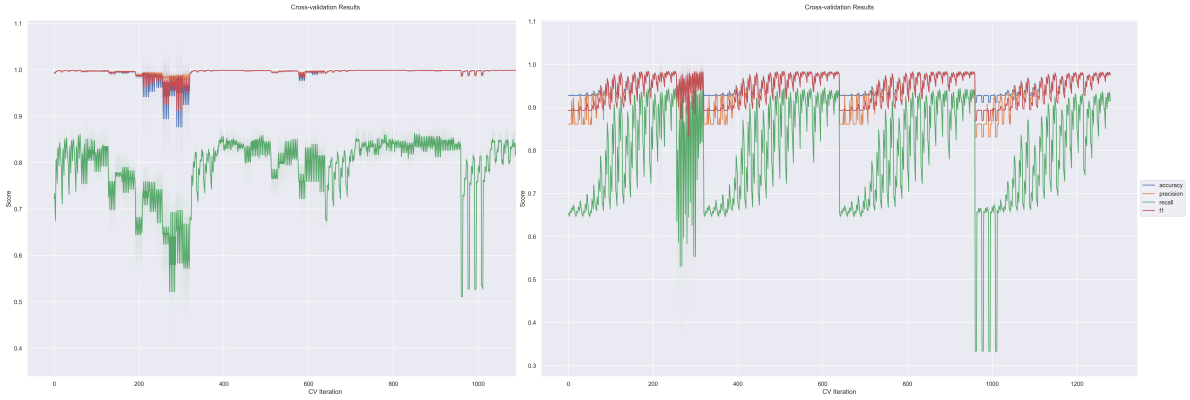


(a) WUSTL-HDRL-2024 dataset (b) WUSTL-EHMS-2020 dataset

Fig. 8: MAE values across two datasets for binary classification

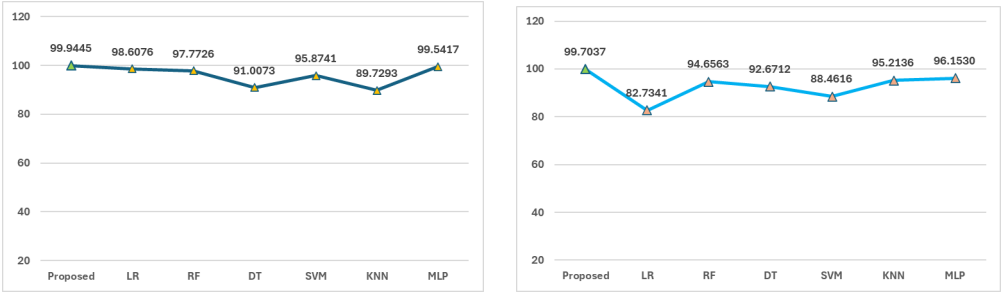
2.1.2 Multi-class classification

For the D2 dataset, SHIELD maintained superior performance with an accuracy of 98.2264%, a precision of 98.1806%, a recall of 93.5687%, and an F1-score of 98.1689% with notable improvement in cross-validation results as shown in Fig. 9. Furthermore, when considering the ROC_AUC metric (Fig. 10), SHIELD achieved values of 99.9445 and 99.7037 for the D1 and D2 datasets, respectively, outperforming the baseline models. Additionally, the MAE as seen in Fig.12 was considerably lower for the proposed method (0.0025 and 0.0181), indicating that its predictions closely match the true values. Although the training times varied, with SHIELD requiring 1.8085 seconds (Fig. 11a) and 2.3150 seconds (Fig. 11b) on the two datasets, respectively, this slight increase in computational time is justified by the substantial improvements in predictive performance.



(a) WUSTL-HDRL-2024 dataset (b) WUSTL-EHMS-2020 dataset

Fig. 9: CV across two datasets for multi-class classification

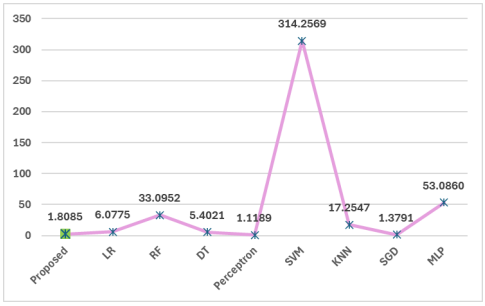


(a) WUSTL-HDRL-2024 dataset (b) WUSTL-EHMS-2020 dataset

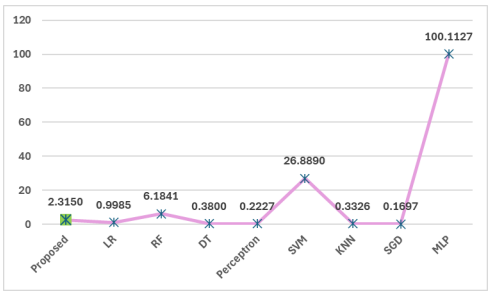
Fig. 10: ROC values across two datasets for multi-class classification

2.2 Discussion

The results highlight SHIELD's consistent and superior performance across diverse datasets and evaluation metrics, confirming its robustness and adaptability for intrusion detection in IoMT environments. The model's exceptionally high precision (99.90%) and F1-score demonstrate its reliability in correctly identifying attacks while maintaining a minimal false-positive rate—an essential attribute in medical contexts where erroneous alerts could disrupt critical operations. Moreover, its outstanding ROC-AUC values indicate strong discriminative power between normal and malicious classes, reflecting its capacity to generalise effectively across varying data distributions. Despite certain baseline methods, such as decision trees and perceptron classifiers, exhibiting faster training times, their overall predictive performance lagged behind SHIELD. The integration of SafeML and Adaptive Gradient Relevance Modelling (AGRM) allows SHIELD to achieve a balanced compromise between computational

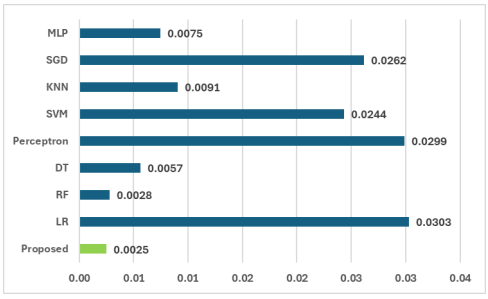


(a) WUSTL-HDRL-2024 dataset

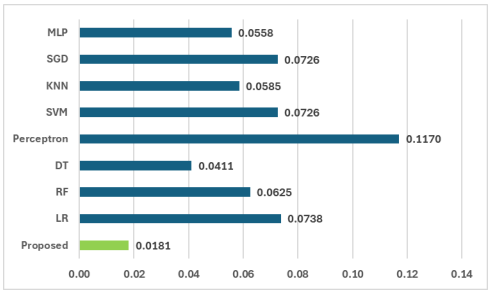


(b) WUSTL-EHMS-2020 dataset

Fig. 11: Average time comparison of various models across two datasets for multi-class classification



(a) WUSTL-HDRL-2024 dataset



(b) WUSTL-EHMS-2020 dataset

Fig. 12: MAE values across two datasets for multi-class classification

efficiency and predictive accuracy, ensuring that its deployment remains feasible even within resource-constrained healthcare infrastructures.

In multi-class classification tasks, SHIELD's consistently higher recall (83.67–93.57%) compared to competing models demonstrates its sensitivity to minority classes—an often-overlooked challenge in healthcare intrusion detection. While models such as Random Forest and Decision Tree attained commendable accuracy, their lower recall (81.57–82.32%) indicated weaker performance in identifying less frequent attack types. The model's minimal Mean Absolute Error (0.0025–0.0181) reinforces its accuracy and stability, positioning it as a reliable tool for real-time cybersecurity monitoring. Furthermore, SHIELD achieved near state-of-the-art results with significantly lower training times than computationally intensive models like SVM and MLP, underscoring its practical efficiency. Although the Random Forest classifier achieved a marginally higher accuracy on one dataset (99.88% vs. 99.90%), SHIELD's overall balance of speed, precision, and adaptability renders it more suitable for scalable, real-world applications. Future enhancements could explore hybrid

ensemble configurations that retain SHIELD's efficiency while further improving resilience against diverse and evolving threat landscapes.

3 Conclusion and Future Work

The integration of IoMT devices into healthcare systems has introduced unparalleled advancements in patient care but also exposed critical vulnerabilities to cyberattacks. Traditional security mechanisms are inadequate for the dynamic and resource-constrained nature of medical devices. This paper addresses these challenges through SHIELD. SHIELD introduces SafeML in integration with Adaptive Gradient Relevance Modelling (AGRM) for optimised feature selection, enhancing detection accuracy while reducing computational overhead. By leveraging real-time pattern learning, SafeML dynamically adapts to evolving threats, achieving 98% accuracy on healthcare intrusion datasets with minimal false positives. Extensive evaluations demonstrate its superiority over conventional models (e.g., RF, SVM, MLP) in precision (99.90% for multi-class), recall (93.57% for multi-class), and ROC-AUC (99.70% for multi-class), alongside efficient training times (1.81–2.32 seconds for multi-class). Our extensive experimental evaluation confirms that the proposed approach not only enhances detection accuracy but also reduces computational requirements, making it particularly well-suited for resource-constrained healthcare settings. The promising results underscore the feasibility of deploying ML-based security solutions in the rapidly evolving landscape of IoT-enabled healthcare. While the model excels on the tested datasets, its performance on highly sparse or noisy data remains unverified. Future work will explore further optimisation of the feature selection process and will focus on validating its scalability and real-time applicability in dynamic environments.