Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Classification of Electromagnetic Side-Channel Activity in 100BASE-TX Ethernet Using a Random Forest Model

ALF Sajeetha

sajeetha92@gmail.com

University of Colombo https://orcid.org/0009-0005-7091-3002

Fathima Shahdhiya

shahdhiya.aj@gmail.com

British College of Applied Studies

Research Article

Keywords: Electromagnetic side-channel analysis, Ethernet, Random Forest, machine learning, signal classification, software-defined radio, feature extraction, ROC-AUC

Posted Date: November 26th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11882790/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: The authors declare no competing interests.

Classification of Electromagnetic Side-Channel Activity in 100BASE-TX Ethernet Using a Random Forest Model

***Abstract*—Electromagnetic emissions from Ethernet cables can provide observable information about network activity without requiring direct electrical access to the communication channel. This study investigates the classification of electromagnetic side-channel activity associated with three operating conditions in a 100BASE-TX Ethernet environment: idle network activity, text-file transfer, and image-file transfer. Electromagnetic signals were captured non-invasively using a HackRF One software-defined radio and an H-field probe positioned near a Cat5e Ethernet cable. The captured complex IQ data were processed offline using overlapping windows of 65,536 samples with 50% overlap. Six signal features were extracted: mean magnitude, root mean square (RMS), standard deviation, peak magnitude, spectral centroid, and spectral bandwidth. A labelled dataset containing 857,130 observations was used to train and evaluate a Random Forest classifier. The optimized model achieved 93.25% accuracy using a stratified random train-test split, while five-fold stratified cross-validation produced a mean accuracy of 93.21%. Class-specific ROC-AUC values ranged from 0.9847 to 0.9931. A separate time-ordered evaluation of the baseline model achieved 91.66% accuracy. Feature analysis identified spectral bandwidth and spectral centroid as the two largest contributors to the optimized model's impurity-based feature importance. However, the evaluation results must be interpreted cautiously because the windows overlap and the training and testing subsets originate from the same original recordings. The study demonstrates activity classification within the collected dataset, but does not establish Ethernet frame decoding, payload reconstruction, independent-session generalization, or real-time classification.**

Keywords—Electromagnetic side-channel analysis, Ethernet, Random Forest, machine learning, signal classification, software-defined radio, feature extraction, ROC-AUC.

I. Introduction

Ethernet is widely used for wired communication in computer networks. Although communication occurs through physical cables, electromagnetic emissions associated with data transmission may provide observable information about network activity. These emissions can be investigated using non-invasive electromagnetic measurement techniques without directly connecting to the electrical communication channel. Consequently, electromagnetic side-channel analysis has potential applications in network security monitoring and the investigation of unintended information leakage [1], [2].

Previous research has investigated electromagnetic emissions from wired communication systems for covert data transmission and side-channel monitoring. Other studies have explored machine learning techniques for classifying electromagnetic activity under different operating conditions. However, distinguishing between electromagnetic activity associated with different Ethernet traffic conditions remains challenging because captured signals may contain noise, background emissions, and variations related to the measurement environment. The acquisition configuration and signal-processing approach can also influence the features available for analysis [1], [2], [3], [4], [5].

This study investigates whether machine learning can distinguish electromagnetic side-channel activity associated with three Ethernet operating conditions: idle network activity, text-file transfer, and image-file transfer. A HackRF One software-defined radio and an H-field probe were used to capture electromagnetic signals near a Cat5e Ethernet cable in a 100BASE-TX network environment. The captured signals were processed offline, and statistical and frequency-domain features were extracted from overlapping signal windows. A Random Forest classifier was then trained and evaluated to determine its ability to distinguish the three activity classes.

The main contributions of this study are the development of a feature-based electromagnetic activity classification pipeline, the evaluation of a Random Forest model using multiple data-splitting strategies, and an examination of the limitations affecting interpretation of the results. The optimized classifier achieved approximately 93.25% accuracy under a stratified random train-test split, while a time-ordered split achieved approximately 91.66%. These findings indicate that the extracted features contain information useful for distinguishing the investigated classes. However, the study focuses on activity classification and does not demonstrate successful Ethernet frame decoding or recovery of transmitted payloads.

II. RELATED WORK

A. Electromagnetic Side-Channel Analysis in Wired Communication

Electromagnetic side-channel analysis investigates unintended electromagnetic emissions produced during the operation of electronic systems. These emissions may reveal information about internal processing or communication activity without requiring direct electrical access to the system. In wired communication environments, Ethernet cables can produce measurable electromagnetic radiation associated with network operation. Sachintha et al. [1] investigated electromagnetic covert-channel data exfiltration through wired industrial control systems, demonstrating the security implications of unintended emissions from Ethernet infrastructure. Similarly, Guri [2] introduced LANTENNA, an electromagnetic exfiltration technique that uses emissions from Ethernet cables to transmit information from air-gapped systems.

More recent work has investigated the use of electromagnetic emissions for network security monitoring. Weerasinghe et al. [3] proposed a non-invasive electromagnetic monitoring approach for forensic-ready industrial control systems and demonstrated the potential to identify selected denial-of-service attack conditions through electromagnetic emission patterns. These studies establish that electromagnetic measurements near wired network

infrastructure can provide useful information for security-related investigations. However, their primary objectives differ from the classification of ordinary Ethernet activity conditions investigated in this research.

B. Side-Channel Measurement and Signal Acquisition

The reliability of side-channel analysis depends on the measurement configuration, sampling process and signal characteristics. Gwinn et al. [4] examined the influence of acquisition and collection parameters on side-channel disassembly, highlighting the importance of measurement configuration and sampling conditions. Their work provides relevant methodological context for considering the limitations of electromagnetic signal acquisition.

The IEEE 802.3 Ethernet standard specifies the physical-layer requirements for Ethernet technologies, including 100BASE-TX operation [5]. Understanding the physical-layer characteristics is important when interpreting electromagnetic measurements collected near Ethernet cables. In particular, the nominal symbol rate of 100BASE-X is substantially higher than the sampling rate used in the present investigation. Therefore, the recorded measurements are analysed as electromagnetic activity observations rather than direct samples of every transmitted Ethernet symbol.

C. Signal Processing and Feature-Based Classification

Digital signal processing provides methods for representing and analysing signals in both time and frequency domains. Welch [6] introduced a periodogram-averaging method for power spectral density estimation, while Harris [7] discussed the use of data windows in discrete Fourier transform-based spectral analysis. Allen [8] and Allen and Rabiner [9] further developed the theoretical basis of short-time spectral analysis. These methods provide a foundation for examining signal magnitude and frequency-distribution characteristics.

Machine learning can use extracted signal features to identify patterns associated with different operating conditions. Breiman [10] introduced the Random Forest algorithm, which combines multiple decision trees to produce an ensemble prediction. Ho [11] examined random decision forests, while Breiman [12] introduced bootstrap aggregation as a method for improving the stability of prediction models. Liaw and Wiener [13] described practical Random Forest classification and regression implementation. Pedregosa et al. [14] presented the scikit-learn machine learning library, which provides tools for model training, parameter optimisation and evaluation.

Electromagnetic side-channel machine learning has also been explored in other security contexts. Welagedara [15] investigated electromagnetic side-channel analysis for fileless malware detection using several machine learning algorithms. Although the research focused on processor emissions and malware-related activity, it demonstrates the broader applicability of combining electromagnetic signal acquisition, feature extraction and machine learning.

D. Research Gap and Motivation

The reviewed studies demonstrate the potential of electromagnetic emissions for information leakage analysis, network monitoring and security-related classification. However, the reviewed works primarily focus on covert data transmission, industrial control system monitoring, malware detection or side-channel disassembly. The classification of ordinary idle, text-file transfer and image-file transfer activity in a controlled 100BASE-TX Ethernet environment represents a different experimental objective.

This research addresses that specific classification problem by extracting statistical and frequency-domain features from offline electromagnetic recordings and evaluating a Random Forest classifier using random and time-ordered data-splitting strategies. The study also examines feature importance, class-specific ROC-AUC values and learning-curve behaviour. Unlike work focused on payload recovery, this investigation evaluates activity classification and does not claim successful Ethernet frame decoding or transmitted-data reconstruction.

III. METHODOLOGY

A. Experimental Setup

An experimental and quantitative approach was adopted to investigate whether electromagnetic measurements collected near an Ethernet cable could distinguish different network activity conditions. The setup consisted of an HP sender computer, a Dell receiver computer, a direct 100BASE-TX Ethernet connection, a HackRF One software-defined radio, and an H-field probe. The sender and receiver used the static IP addresses 192.168.1.1 and 192.168.1.2, respectively. The Ethernet link operated at a nominal data rate of 100 Mbps. The probe was positioned near the cable without direct electrical contact, allowing electromagnetic signals associated with network activity to be recorded non-invasively. The experimental arrangement used for electromagnetic signal acquisition is illustrated in Fig. 1.

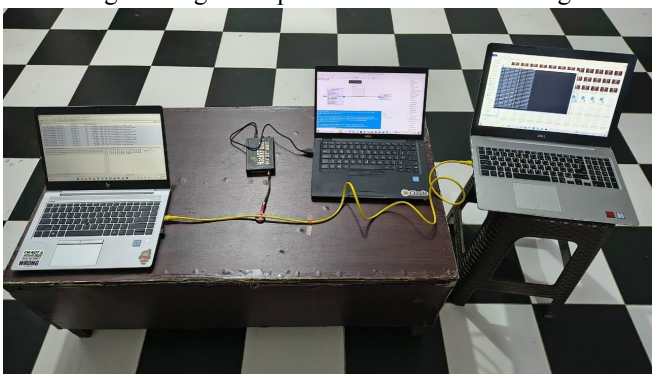

Figure 1 Experimental setup for non-invasive electromagnetic signal acquisition from a 100BASE-TX Ethernet cable.

B. Electromagnetic Signal Acquisition

Electromagnetic signals were captured using the HackRF One at a centre frequency of 100 MHz and a sampling rate of 10.5 MS/s. The recorded configuration used RF, IF, and baseband gains of 10, 20, and 20 dB, respectively. The approximate captured frequency span was 94.75–105.25 MHz. Signals were stored as complex float32 in-phase and quadrature (IQ) samples in CFile format.

Three experimental conditions were recorded: idle network activity, text-file transfer, and image-file transfer. The recordings were processed offline because the investigation used stored signal captures rather than a real-time

classification system. Corresponding PCAPNG files were collected using network packet-capture tools to provide separate evidence of the traffic generated during the experiments.

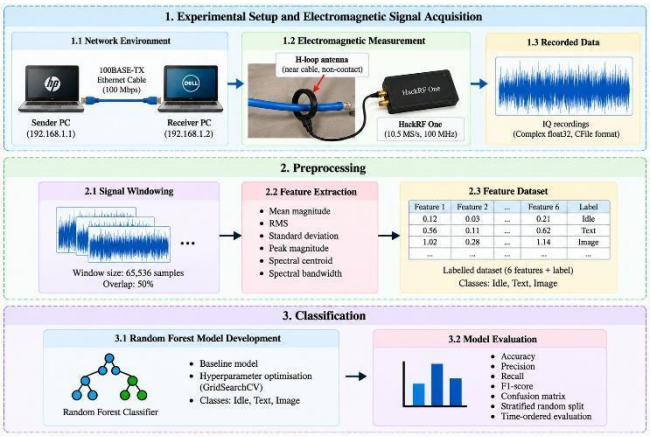

Figure 2 High-level architecture of the experimental setup, electromagnetic signal acquisition, preprocessing, and Random Forest classification workflow.

C. *Dataset Preparation and Feature Extraction*

The recorded IQ data were divided into fixed-size windows containing 65,536 samples, with a 50% overlap between consecutive windows. Six features were extracted from each window: mean magnitude, root mean square (RMS), standard deviation, peak magnitude, spectral centroid, and spectral bandwidth. These features represented signal magnitude, variability, and frequency-distribution characteristics.

The resulting labelled dataset contained 857,130 windows and eight columns, comprising a segment identifier, six numerical features, and a class label. The three classes were Idle, Text, and Image. The dataset was used as input to the supervised classification pipeline. Packet-capture statistics were analysed separately to investigate possible relationships between observed network traffic and electromagnetic measurements.

D. *Random Forest Model Development*

A Random Forest classifier was selected to classify the extracted feature vectors into the three activity categories. A baseline model was evaluated before hyperparameter optimisation was performed using GridSearchCV with five-fold cross-validation. The optimisation process considered different combinations of tree count, maximum tree depth, minimum samples per leaf, and minimum samples required for a split.

The best-performing configuration contained 200 trees, a maximum depth of 20, a minimum of two samples per leaf, and a minimum of five samples required for a split. The resulting model was evaluated using the extracted features and their corresponding activity labels.

E. *Model Evaluation*

Model performance was assessed using accuracy, precision, recall, F1-score, and confusion-matrix analysis. A stratified random 80:20 train-test split was used to evaluate classification performance while maintaining class proportions. Five-fold stratified cross-validation was also conducted to assess performance across multiple folds.

In addition, a time-ordered evaluation trained the model on earlier windows and tested it on later windows. This provided a complementary assessment of performance when the data were separated chronologically. However, both evaluation approaches used windows from the same original recordings; therefore, the time-ordered split did not establish generalisation to independent recording sessions. The evaluation focused on classifying activity categories rather than decoding Ethernet frames or reconstructing transmitted payloads.

The complete processing and classification procedure is summarised in Algorithm 1. Overlapping windows of 65,536 IQ samples are generated from each recording, six statistical and spectral features are extracted from each window, and a Random Forest classifier is trained and evaluated using stratified random, cross-validated and time-ordered strategies.

Algorithm 1. Electromagnetic side-channel activity classification pipeline.

Input: IQ recordings R_c for each class c ∈ {Idle, Text, Image}; window length L = 65,536; hop H = L/2 (50% overlap); sampling rate fs = 10.5 MS/s
Output: trained Random Forest M*, performance metrics

D ← ∅
for each class c ∈ {Idle, Text, Image} **do**
  Load complex IQ samples x[n] from R_c (CFile, float32)
  for i = 0, H, 2H, … while i + L ≤ length(x) **do**
    w ← x[i : i + L]
    m ← |w| ▷ magnitude
    f1 ← mean(m) ▷ mean magnitude
    f2 ← sqrt(mean(m²)) ▷ RMS
    f3 ← std(m) ▷ standard deviation
    f4 ← max(m) ▷ peak magnitude
    P ← |FFT(w)|; freq ← FFT frequency bins
    f5 ← Σ(freq · P) / Σ P ▷ spectral centroid
    f6 ← sqrt(Σ((freq − f5)² · P) / Σ P) ▷ spectral bandwidth
    Append (segment_id, f1…f6, label = c) to D
  end for
end for
Split D into train/test (80:20), stratified by class
Baseline: train Random Forest M_b on the training set; evaluate on the test set
Tune with GridSearchCV (5-fold) over n_trees, max_depth, min_samples_leaf, min_samples_split
M* ← best model (200 trees, depth 20, leaf 2, split 5)
Evaluate M* on the test set: accuracy, precision, recall, F1-score, confusion matrix, one-vs-rest ROC-AUC per class
Run five-fold stratified cross-validation; report mean ± std accuracy
Time-ordered evaluation: train on earlier windows, test on later windows
Compute impurity-based feature importance of M*
return M*, metrics

IV. RESULTS AND DISCUSSION

A. *Dataset and Feature Statistics*

The prepared dataset contained 857,130 labelled signal-window observations distributed across three network activity conditions: Idle, Text-file transfer and Image-file transfer. The Idle class contained 293,415 observations, the Text class contained 282,735 observations, and the Image class contained 280,980 observations. The class distribution was therefore relatively balanced, with each class

contributing approximately one-third of the total observations.

The six extracted features represented signal magnitude, variation and frequency-domain characteristics. Descriptive statistics were calculated separately for each activity class to examine the numerical behavior of the extracted variables. The mean magnitude values were 0.2335 for Idle, 0.1595 for Text and 0.1991 for Image. The corresponding RMS values were 0.2581, 0.1795 and 0.2276. These results show differences in the average magnitude-related feature values across the three recorded activity conditions.

The mean standard deviation values were approximately 0.1100 for Idle, 0.0822 for Text and 0.1101 for Image. The mean spectral centroid values were approximately 2.345 MHz, 2.312 MHz and 2.290 MHz for Idle, Text and Image, respectively. The mean spectral bandwidth values were approximately 1.419 MHz for Idle, 1.408 MHz for Text and 1.348 MHz for Image.

These descriptive results indicate that the recorded activity classes exhibited differences in several extracted signal features. However, the descriptive statistics alone do not establish that the differences are statistically significant. Instead, the classification experiments examine whether the combined feature information can distinguish the labelled activity conditions.

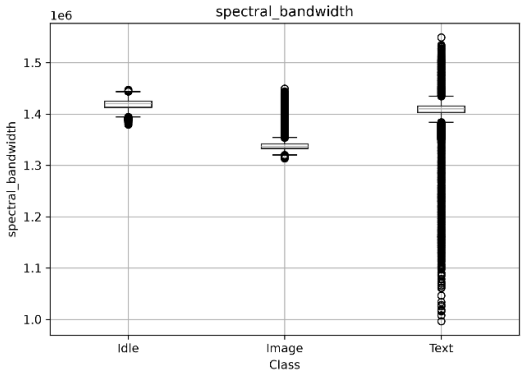

Figure 3 Boxplot distribution of spectral bandwidth across Idle, Text, and Image activity classes.

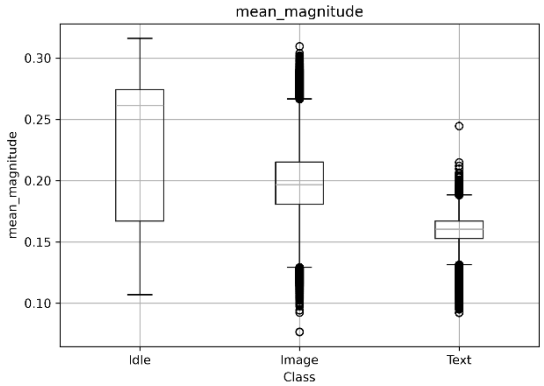

Figure 4 Boxplot distribution of mean magnitude across Idle, Text, and Image activity classes.

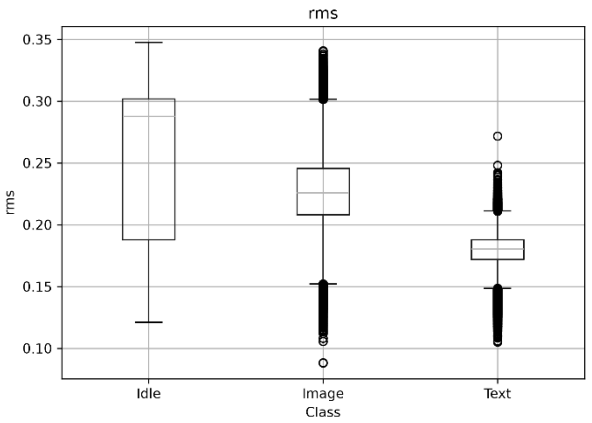

Figure 5 Boxplot distribution of root mean square (RMS) values across Idle, Text, and Image activity classes.

B. Random Forest Classification Performance

The baseline Random Forest classifier achieved an accuracy of 93.20% using a stratified random 80:20 train-test split. Following hyperparameter optimization, the best configuration contained 200 decision trees, a maximum depth of 20, a minimum of two samples per leaf and a minimum of five samples required to split an internal node.

The optimized model achieved an accuracy of approximately 93.25% on the random test subset. The difference between the baseline and optimized models was approximately 0.05 percentage points, indicating a modest improvement following hyperparameter optimization.

Five-fold stratified cross-validation produced fold accuracies of approximately 93.15%, 93.24%, 93.26%, 93.17% and 93.23%. The mean accuracy was 93.21%, with a standard deviation of approximately 0.04 percentage points. These results indicate that the observed accuracy remained relatively consistent across the five validation folds under the selected evaluation procedure.

The class-level classification report showed that the Image class achieved a precision of 0.97, recall of 0.94 and F1-score of 0.95. The Idle class achieved a precision of 0.94, recall of 0.93 and F1-score of 0.94. The Text class achieved a precision of 0.89, recall of 0.93 and F1-score of 0.91. The lower precision and F1-score for Text indicate that this class was comparatively more difficult for the model to distinguish correctly.

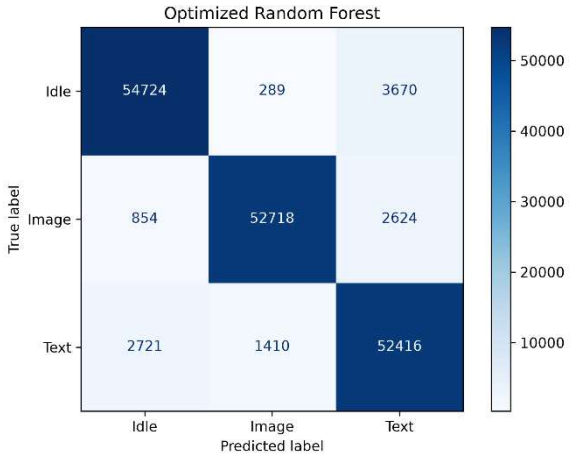

Figure 6 Confusion matrix of the baseline Random Forest classifier for the three Ethernet activity classes.

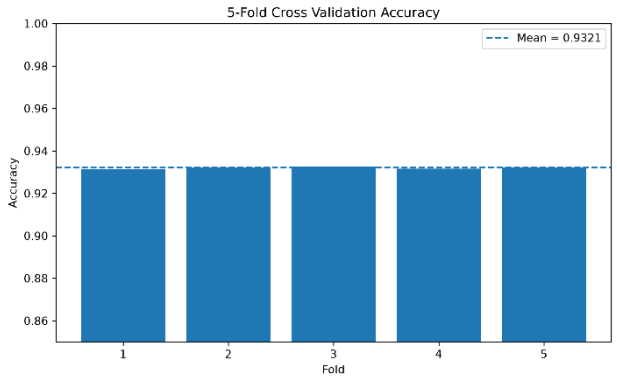

Figure 7 Classification accuracy obtained across the five stratified cross-validation folds.

C. *Feature Correlation and Importance Analysis*

The feature correlation heatmap was examined to investigate relationships among the extracted numerical variables. The heatmap showed a very strong positive correlation between mean magnitude and RMS, with a reported correlation of approximately 1.00. Standard deviation was also strongly correlated with mean magnitude and RMS, with reported coefficients of approximately 0.93 and 0.95, respectively. Spectral centroid and spectral bandwidth showed a moderate positive correlation of approximately 0.59.

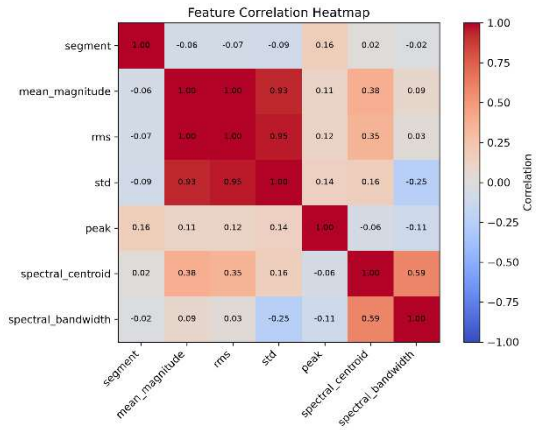

Figure 8 Correlation matrix illustrating relationships among the extracted electromagnetic signal features.

These relationships indicate that some of the extracted features describe related aspects of the recorded signals. In particular, mean magnitude and RMS provide closely related information about signal magnitude. The correlation results should therefore be considered when interpreting the contribution of individual features to classification performance.

The optimized Random Forest feature-importance output identified spectral bandwidth as the largest contributor, with an importance value of approximately 0.2668. Spectral centroid followed with 0.2414, while standard deviation contributed approximately 0.1937. Mean magnitude and RMS contributed approximately 0.1562 and 0.1420, respectively.

Peak magnitude received an importance value of zero. Examination of the feature summary statistics showed that the peak feature had a value of approximately 1.0 across all three classes, with an almost zero standard deviation. This indicates that the feature was effectively constant in the prepared dataset and therefore provided little or no discriminatory information to the classifier. This observation highlights the importance of validating extracted features before model training.

The feature-importance values represent the model's impurity-based importance estimates and should not be interpreted as proof of causal influence. Furthermore, the strong correlation between some magnitude-related features means that importance may be distributed across related variables.

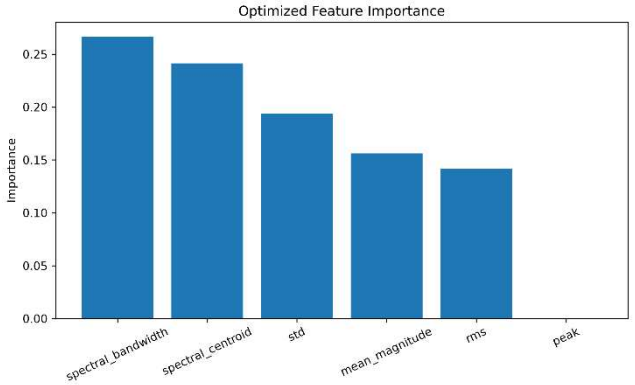

Figure 9 Impurity-based feature importance scores of the optimized Random Forest classifier.

D. *ROC-AUC Analysis*

The classification model was also examined using class-specific Receiver Operating Characteristic Area Under the Curve (ROC-AUC) results. The recorded AUC values were 0.9913 for Idle, 0.9931 for Image and 0.9847 for Text.

Among the three classes, Image produced the highest recorded AUC value, followed by Idle and Text. These values indicate strong separation between each class and the remaining classes under the ROC evaluation procedure used in the experiment. The lower AUC value for Text is consistent with the comparatively lower F1-score reported for that class.

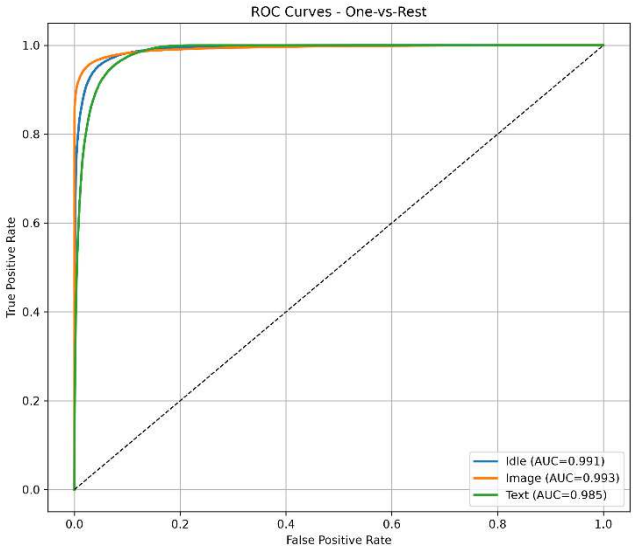

Figure 10 Receiver Operating Characteristic (ROC) curves for the Idle, Text, and Image activity classes.

ROC-AUC provides an additional perspective on classification performance because it examines the relationship between true-positive and false-positive rates across classification thresholds. However, these results are derived from the same experimental dataset and should not be interpreted as independent-session validation. The ROC-AUC results therefore complement, rather than replace, the accuracy, precision, recall, F1-score and confusion-matrix analysis.

E. Learning Curve Analysis

The learning curve was examined to investigate how training and validation accuracy changed as the number of training observations increased. The saved results showed that training accuracy decreased from approximately 97.92% at 16,000 training observations to approximately 96.87% at 160,000 observations. In contrast, validation accuracy increased from approximately 91.59% to approximately 92.97% over the same range.

The reduction in training accuracy and increase in validation accuracy suggest that the model's generalization estimate improved as more training observations were included. At the final recorded training size, a difference remained between training accuracy and validation accuracy. This indicates that the model continued to fit the training data more closely than the validation data.

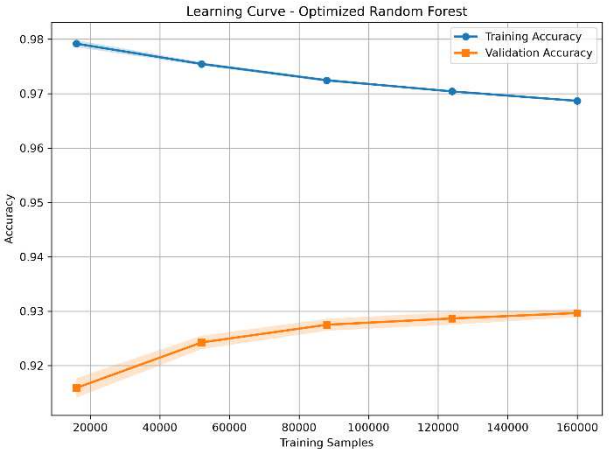

Figure 11 Training and validation accuracy of the Random Forest classifier with increasing training-set size.

The learning curve provides additional information about the relationship between training-set size and predictive performance. Nevertheless, it does not resolve the potential dependency introduced by overlapping windows or establish performance on independently collected recording sessions.

F. Time-Frequency and Frequency-Band Analysis

Short-time Fourier transform (STFT) visualizations were generated for the Idle, Text and Image activity recordings to examine their time-frequency characteristics. These visualizations provide an additional way to inspect how spectral content varies over the recorded intervals. They complement the numerical feature-extraction process by presenting the distribution of signal energy across time and frequency.

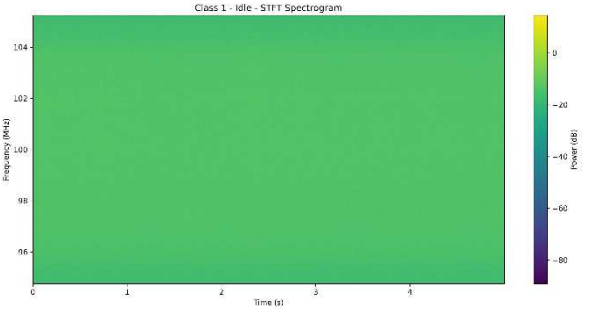

Figure 12 Short-Time Fourier Transform (STFT) representation of the Idle activity recording.

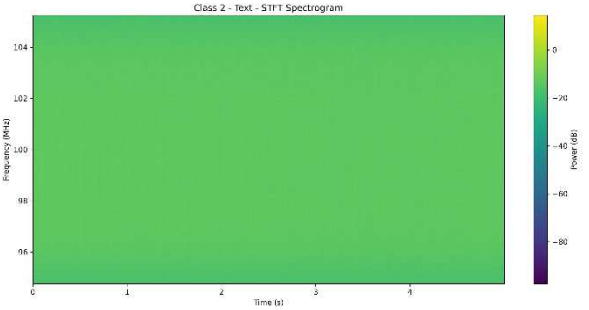

Figure 13 Short-Time Fourier Transform (STFT) representation of the Text-file transfer recording.

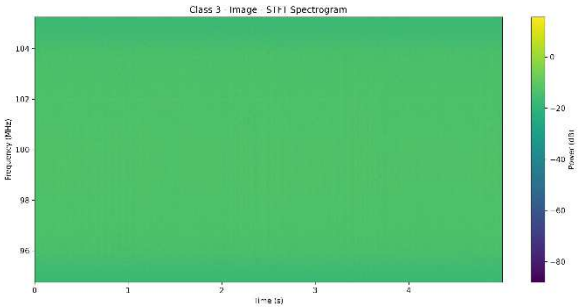

Figure 14 Short-Time Fourier Transform (STFT) representation of the Image-file transfer recording.

The frequency-band analysis also examined the recorded peak frequency and peak power at selected time intervals. The preprocessing summary showed that the peak frequency for the Idle and Text recordings remained close to 101.996 MHz across the reported measurement points. The Image recording showed more variation in its recorded peak frequency, including values near 101.869 MHz, 101.670 MHz and 98.983 MHz at selected intervals.

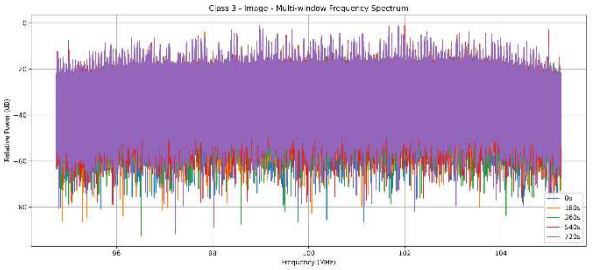

Figure 15 Frequency-band analysis of the electromagnetic signal recorded during Image-file transfer.

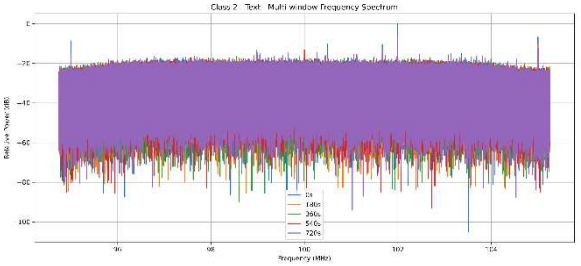

Figure 16 Frequency-band analysis of the electromagnetic signal recorded during Text-file transfer.

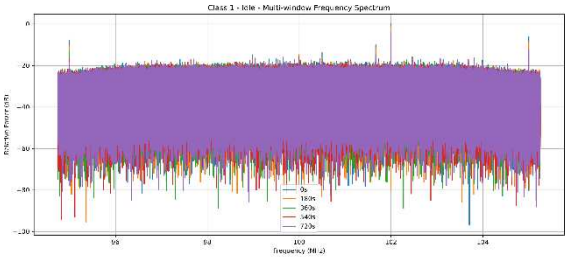

Figure 17 Frequency-band analysis of the electromagnetic signal recorded during Idle network activity.

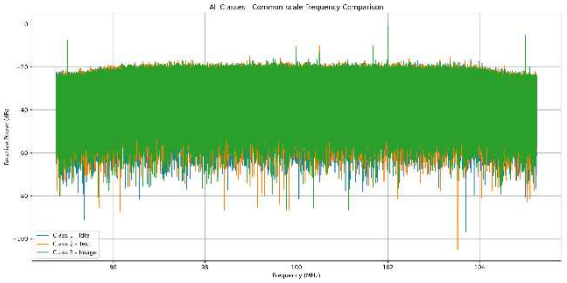

Figure 18 Comparison of electromagnetic frequency-band characteristics across Idle, Text, and Image activity classes**.**

The recorded peak-power values also varied over time and between activity classes. These observations demonstrate that the captured electromagnetic recordings contained time-dependent spectral characteristics. However, the plots and summary values do not independently establish that a particular spectral variation was caused exclusively by a specific file-transfer activity. Further controlled experiments and independent recordings would be needed to establish the repeatability of these patterns.

G. Overall Interpretation

The combined results indicate that the extracted electromagnetic signal features contained information that enabled the Random Forest model to distinguish the three selected activity conditions within the prepared dataset. The optimized random-split accuracy was approximately 93.25%, five-fold cross-validation produced a mean accuracy of 93.21%, and the class-specific ROC-AUC values ranged from 0.9847 to 0.9931.

The feature analysis identified spectral bandwidth and spectral centroid as the most important variables according to the model's impurity-based importance estimates. The analysis also identified peak magnitude as a constant feature in the prepared dataset, explaining its zero importance. The learning curve and time-ordered evaluation provided further information about model behavior and the influence of evaluation strategy.

Despite these results, the use of overlapping windows means that related observations may appear in both training and testing subsets. The time-ordered evaluation also used recordings from the same original experimental sessions. Consequently, the reported metrics demonstrate classification performance within the collected dataset but do not establish generalization to independently recorded sessions.

The study therefore supports the feasibility of electromagnetic activity classification for the three investigated Ethernet conditions. It does not demonstrate complete Ethernet frame decoding, transmitted-payload reconstruction, reliable network traffic-volume estimation or real-time classification.

V. LIMITATIONS

Several limitations should be considered when interpreting the results of this study. First, electromagnetic signals were captured at a sampling rate of 10.5 MS/s, which is substantially lower than the nominal 125 MBd symbol rate of 100BASE-TX Ethernet. Therefore, the captured signals do not provide sufficient temporal sampling to directly

observe every transmitted symbol, and successful Ethernet frame decoding or payload reconstruction was not demonstrated.

Second, the feature windows overlapped by 50%. Consequently, randomly dividing the dataset into training and testing sets may place similar signal segments in both sets, potentially producing optimistic performance estimates. Although the time-ordered evaluation achieved 91.66% accuracy, it used data from the same original recordings and therefore does not establish generalisation to independent recording sessions.

Third, the experiments were conducted using a single Cat5e Ethernet cable and a controlled two-computer setup. Only three activity classes—idle, text-file transfer, and image-file transfer—were investigated. The model has not been validated across different cables, hardware configurations, environments, or independently collected datasets.

Finally, the processing and classification were performed offline rather than in real time. The observed correlations between electromagnetic measurements and packet-capture statistics were weak over the full recording durations. Therefore, the findings support classification of the investigated activity conditions within the collected dataset, but do not establish reliable traffic-volume estimation, real-time performance, or recovery of transmitted data.

An additional limitation was identified during feature validation. The peak-magnitude feature had an approximately constant value of 1.0 across all three activity classes and consequently received zero impurity-based feature importance in the optimised Random Forest model. This suggests that the feature-extraction or normalisation procedure did not produce useful variation for this feature in the prepared dataset. In addition, the strong correlations between mean magnitude and RMS indicate redundancy among some of the extracted variables. These observations highlight the need for further feature validation and selection in future experiments.

VI. CONCLUSION AND FUTURE WORK

This study investigated the classification of electromagnetic side-channel activity associated with idle network conditions, text-file transfers and image-file transfers in a 100BASE-TX Ethernet environment. Electromagnetic signals were captured non-invasively using a HackRF One software-defined radio and an H-field probe positioned near a Cat5e Ethernet cable. The captured IQ data were processed offline, and six numerical features were extracted from overlapping signal windows to construct a labelled dataset containing 857,130 observations.

A Random Forest classifier was developed and evaluated using stratified random train-test splitting, five-fold cross-validation and a time-ordered evaluation. The optimized model achieved approximately 93.25% accuracy under the random-split evaluation, while five-fold cross-validation produced a mean accuracy of 93.21%. The baseline model achieved approximately 91.66% accuracy under the time-ordered evaluation. The class-specific ROC-AUC values ranged from 0.9847 to 0.9931, providing additional evidence of class separation within the evaluated dataset.

Feature analysis identified spectral bandwidth and spectral centroid as the largest contributors to the optimized model's impurity-based feature importance. The analysis also identified a constant peak-magnitude feature with zero importance, demonstrating the value of examining extracted features rather than relying exclusively on overall classification accuracy. Learning-curve results showed increasing validation accuracy as the training-set size increased, although a difference between training and validation accuracy remained.

The findings indicate that the extracted electromagnetic signal features contain information useful for distinguishing the three investigated Ethernet activity conditions within the collected dataset. However, the use of overlapping windows, the absence of independent recording-session validation and the limited experimental configuration must be considered when interpreting the results. The study does not demonstrate successful Ethernet frame decoding, transmitted-payload reconstruction, reliable traffic-volume estimation or real-time classification.

Future work should collect independent recording sessions and evaluate the classifier using session-separated training and testing datasets. Additional research should investigate the peak-feature extraction issue, reduce unnecessary feature redundancy, examine alternative signal representations and compare suitable classification algorithms under consistent evaluation conditions. Experiments involving different Ethernet cables, probe positions, hardware configurations and network activity classes would also help determine the generalizability of the approach. Finally, further hardware and software development would be required to investigate the feasibility of continuous real-time electromagnetic activity classification.

In conclusion, the research establishes an experimental feature-based approach for classifying three controlled Ethernet activity conditions using electromagnetic measurements and a Random Forest model. The results provide a basis for further investigation into non-invasive electromagnetic activity classification in wired communication environments.

VII. BIBLIOGRAPHY

[1] S. Sachintha, N.-A. Le-Khac, M. Scanlon and A. P. Sayakkara, "Data exfiltration through electromagnetic covert channel of wired industrial control systems," *Applied Sciences,* vol. 13, no. 5, p. 2928, 2023.

[2] M. Guri, "LANTENNA: Exfiltrating data from air-gapped networks via Ethernet cables," in *IEEE 45th Annual Computers, Software, and Applications Conference (COMPSAC)*, 2021.

[3] B. Weerasinghe, A. Sayakkara, K. D. Zoysa and M. Scanlon, "Low-overhead and non-invasive electromagnetic side-channel monitoring for forensic-ready industrial control systems," in *Digital Forensics Doctoral Symposium (DFDS)*, 2025.

[4] R. Gwinn, M. Matties and A. D. Rubin, "Configuration and collection factors for side-channel disassembly," arXiv, 2022.

[5] A. W. Y. N. Welagedara, "Fileless malware detection using electromagnetic side-channel analysis," Colombo, Sri Lanka, 2025.

[6] IEEE, "IEEE Standard for Ethernet," IEEE, 2022.

[7] L. Breiman, "Random forests," *Machine Learning,* vol. 45, no. 1, p. 5–32, 2001.

[8] T. K. Ho, "Random decision forests," in *3rd International Conference on Document Analysis and Recognition (ICDAR)*, 1995.

[9] L. Breiman, "Bagging predictors," *Machine Learning,* vol. 24, no. 2, p. 123–140, 1996.

[10] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R News,* vol. 2, no. 3, p. 18–22, 2002.

[11] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research,* vol. 12, p. 2825–2830, 2011.

[12] T. Hastie, R. Tibshirani and J. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, 2nd ed., New York, NY, USA: Springer, 2009.

[13] A. V. Oppenheim and R. W. Schafer, Discrete-Time Signal Processing, 3rd ed., Upper Saddle River, NJ, USA: Prentice Hall, 2010.

[14] P. D. Welch, "The use of fast Fourier transform for the estimation of power spectra: A method based on time averaging over short, modified periodograms," *IEEE Transactions on Audio and Electroacoustics,* vol. 15, no. 2, p. 70–73, June 1967.

[15] F. J. Harris, "On the use of windows for harmonic analysis with the discrete Fourier transform," *Proceedings of the IEEE,* vol. 66, no. 1, p. 51–83, January 1978.

[16] J. B. Allen, "Short term spectral analysis, synthesis, and modification by discrete Fourier transform," *IEEE Transactions on Acoustics, Speech, and Signal Processing,* vol. 25, no. 3, p. 235–238, June 1977.

[17] J. B. Allen and L. R. Rabiner, "A unified approach to short-time Fourier analysis and synthesis," *Proceedings of the IEEE,* vol. 65, no. 11, p. 1558–1564, November 1977.

[18] T. Fawcett, "An introduction to ROC analysis," *Pattern Recognition Letters,* vol. 27, no. 8, p. 861–874, 2006.

[19] J. Davis and M. Goadrich, "The relationship between Precision-Recall and ROC curves," in *23rd International Conference on Machine Learning (ICML)*, 2006.

[20] R. Kohavi, "A study of cross-validation and bootstrap for accuracy estimation and model selection," in *4th International Joint Conference on Artificial Intelligence (IJCAI)*, 1995.

[21] G. C. Cawley and N. L. C. Talbot, "On over-fitting in model selection and subsequent selection bias in performance evaluation," *Journal of Machine Learning Research,* vol. 11, p. 2079–2107, 2010.

[22] C. Strobl, A.-L. Boulesteix, A. Zeileis and T. Hothorn, "Bias in random forest variable importance measures: Illustrations, sources and a solution," *BMC Bioinformatics,* vol. 8, p. 25, 2007.

[23] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the ROC plot when evaluating binary classifiers on imbalanced datasets," *PLOS ONE,* vol. 10, no. 3, p. e0118432, 2015.

[24] M. Sokolova and G. Lapalme, "A systematic analysis of performance measures for classification tasks," *Information Processing & Management,* vol. 45, no. 4, p. 427–437, 2009.

[25] J. R. Quinlan, C4.5: Programs for Machine Learning, San Mateo, CA, USA: Morgan Kaufmann, 1993.