IJNRD.ORG ISSN : 2456-4184
INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD) | IJNRD.ORG
An International Open Access, Peer-reviewed, Refereed Journal

MALWARE DETECTION USING MACHINE LEARNING

Chalamalla Nikhil kumar reddy
Student
CMR institute of techonology

Abstract:

Malicious software, commonly referred to as malware, continues to be a persistent and evolving threat in the realm of cybersecurity. As traditional signature-based detection methods struggle to keep pace with the dynamic nature of malware, the integration of machine learning techniques has emerged as a promising avenue for enhancing detection accuracy. This paper offers a comprehensive review and analysis of the current state of malware detection using machine learning.

The review begins by outlining the challenges posed by the ever-changing landscape of malware, emphasizing the limitations of conventional detection methods. The subsequent exploration of machine learning algorithms, including supervised learning (e.g., Decision Trees, Support Vector Machines), unsupervised learning (e.g., Clustering algorithms), and deep learning (e.g., Neural Networks), highlights their potential in improving detection capabilities.

A critical component of effective malware detection is feature extraction, and this paper delves into various static, dynamic, and hybrid analysis features. It underscores the importance of feature selection in refining the accuracy of machine learning models.

The discussion extends to datasets used for training and evaluation, examining publicly available datasets such as the Malware Genome Project and the Microsoft Malware Classification Challenge. The evaluation metrics, including Precision, Recall, and F1 Score, are elucidated, along with the challenges inherent in assessing machine learning models for malware detection.

Furthermore, the paper identifies and discusses challenges and limitations in the application of machine learning, such as adversarial attacks, imbalanced datasets, and the need for models to generalize across rapidly evolving malware variants. These challenges underscore the importance of ongoing research to address these issues.

Looking ahead, the paper outlines future directions in the field, including the incorporation of explainability in machine learning models, the utilization of ensemble learning for improved accuracy, and the exploration of real-time and proactive detection approaches. The integration of machine learning with threat intelligence feeds is also proposed as a promising avenue for enhancing overall cybersecurity.

In conclusion, this paper provides a thorough examination of the use of machine learning in malware detection, offering insights into the current state of the field, its challenges, and future directions. It aims to serve as a valuable resource for researchers, practitioners, and cybersecurity professionals engaged in fortifying defenses against the persistent and adaptive threat of malware.

Introduction:

In the ever-evolving landscape of information technology, the menace of malicious software, or malware, stands as a formidable challenge to the security and integrity of computer systems and networks. As technological advancements progress, so too does the sophistication and diversity of malware, necessitating the continuous evolution of detection mechanisms. Traditional approaches, particularly signature-based

detection methods, face inherent limitations in keeping pace with the rapid mutation and polymorphic nature of contemporary malware.

This introduction seeks to underscore the escalating need for advanced detection techniques and introduces the pivotal role that machine learning plays in addressing this imperative. Malware, encompassing viruses, worms, trojans, ransomware, and more, exhibits a spectrum of behaviors that defy easy categorization through static signatures alone. Consequently, the exploration of machine learning methodologies becomes crucial for their ability to discern complex patterns and anomalies inherent in malware behavior.

The purpose of this paper is to conduct a thorough examination of the integration of machine learning techniques into the realm of malware detection. Through this exploration, we aim to shed light on the efficacy, challenges, and future prospects of employing machine learning to fortify cybersecurity measures against the persistent and adaptive threat posed by malware.

The following sections will delve into the shortcomings of traditional detection methods, providing a foundation for understanding the necessity of machine learning. The paper will subsequently navigate through various machine learning algorithms, emphasizing their potential in augmenting detection accuracy. A critical aspect of this exploration will be the discussion of feature extraction techniques, elucidating how they contribute to the capability of machine learning models to discriminate between benign and malicious software. Furthermore, the review will encompass an analysis of datasets commonly employed for training and evaluation in machine learning-based malware detection, highlighting the importance of representative and diverse datasets. Evaluation metrics and challenges associated with assessing the performance of machine learning models will also be discussed.

This examination is not merely retrospective; it is forward-looking. The paper will scrutinize the challenges inherent in the application of machine learning to malware detection, addressing issues such as adversarial attacks, imbalanced datasets, and the generalization of models across evolving malware variants. In light of these challenges, the paper will conclude with a forward-thinking discussion on potential avenues for future research and development in the field of malware detection using machine learning.

In essence, this introduction sets the stage for an in-depth exploration of the intersection between machine learning and malware detection, emphasizing the urgency of innovative approaches to fortify our cyber defenses against an increasingly sophisticated and diverse array of malicious s

Traditional Malware Detection Methods:

The historical battleground between cybersecurity experts and the persistent threat of malware has witnessed the evolution of several conventional detection methods. These methods, while foundational, face inherent challenges in addressing the dynamic and polymorphic nature of contemporary malicious software.

1. Signature-based Detection:
   - **Principle**: This method involves the creation and utilization of signatures, or unique patterns, associated with known malware. When a file or piece of code matches a predefined signature, it is identified as malicious.
   - **Strengths**: Signature-based detection is efficient and quick, making it suitable for identifying known threats.
   - **Limitations**: The main limitation lies in its inability to detect previously unknown or mutated malware variants, rendering it less effective against the constantly evolving threat landscape.

2. **Heuristic-based Detection**:
   - **Principle**: Heuristic analysis involves the examination of code or file behavior based on predetermined rules and algorithms. Unlike signature-based detection, heuristics focus on identifying suspicious patterns or behaviors that may indicate malware.
   - **Strengths**: This method can detect previously unknown threats by identifying behaviors common to malware.
   - **Limitations**: Heuristics may generate false positives and are dependent on predefined rules, potentially missing novel attack vectors.

3. **Behavior-based Detection**:
   - **Principle**: Behavior-based detection observes the runtime behavior of programs or processes. Deviations from normal behavior, such as unauthorized access or unusual system interactions, trigger alerts.
   - **Strengths**: This method can identify both known and unknown threats based on their behavior during execution.
   - **Limitations**: The challenge lies in distinguishing between malicious and legitimate behaviors accurately. It may also be resource-intensive, impacting system performance.

While these traditional methods have been foundational in mitigating the risks associated with malware, their limitations are increasingly apparent in the face of sophisticated and polymorphic threats. As malware continues to evolve, the necessity for more adaptive and intelligent detection mechanisms becomes apparent. The subsequent sections of this paper will delve into the potential of machine learning as a solution to address these limitations and enhance the overall efficacy of malware detection.

Machine Learning in Malware Detection:

In response to the limitations of traditional detection methods, the integration of machine learning (ML) has emerged as a transformative approach to enhance the accuracy and adaptability of malware detection. Machine learning leverages algorithms and statistical models to enable systems to learn from data, improving their ability to identify patterns, anomalies, and trends. In the context of malware detection, machine learning holds the promise of addressing the challenges posed by the ever-evolving nature of malicious software.

1. **Supervised Learning**:
   - **Overview**: In supervised learning, the model is trained on labeled datasets, where each sample is paired with its corresponding label (malicious or benign). The model learns to make predictions based on this labeled training data.
   - **Algorithms**: Common algorithms include Decision Trees, Support Vector Machines (SVM), and Random Forests. These models excel in classification tasks and can discern intricate patterns within feature-rich datasets.

2. **Unsupervised Learning**:
   - **Overview**: Unsupervised learning operates without labeled training data. The algorithm explores the inherent structure within the data, identifying patterns and anomalies without predefined labels.
   - **Algorithms**: Clustering algorithms, such as K-Means and Hierarchical Clustering, are commonly employed. Unsupervised learning is particularly useful for detecting unknown or novel malware variants.

3. **Deep Learning**:
   - **Overview**: Deep learning employs neural networks with multiple layers (deep neural networks) to automatically learn hierarchical representations of data. It has demonstrated remarkable success in complex pattern recognition tasks.
   - **Algorithms**: Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are often utilized for image and sequence data, respectively. Deep learning excels in feature learning and abstraction, enabling the detection of subtle patterns in malware behavior.

4. Feature Extraction Techniques:
   - Static Analysis Features: Involves examining file attributes, such as file size, entropy, and file header information, without executing the code.
   - Dynamic Analysis Features: Focuses on runtime behavior, observing actions during code execution, system calls, and network activities.
   - Hybrid Analysis Features: Integrates both static and dynamic analysis for a comprehensive understanding of malware behavior.

5. Advantages of Machine Learning:
   - Adaptability: ML models can adapt to new and evolving malware threats by learning from updated datasets.
   - Automation: ML automates the process of feature selection and model training, reducing the need for manual intervention.
   - Scalability: ML can scale to analyze large and diverse datasets efficiently.

While machine learning offers promising solutions for improving malware detection, challenges such as overfitting, adversarial attacks, and the need for diverse and representative datasets persist. The subsequent sections of this paper will delve into these challenges, as well as the evaluation metrics, datasets, and future directions in the realm of machine learning-based malware detection.

Datasets for Training and Evaluation:

The selection of appropriate datasets for training and evaluating machine learning models is a critical aspect of developing robust and effective malware detection systems. Datasets serve as the foundation for training

models to recognize patterns indicative of malicious behavior and subsequently assessing their performance. Several publicly available datasets are widely used in the field of malware detection, each offering unique characteristics and challenges.

1. Malware Genome Project:
- Description: The Malware Genome Project provides a diverse collection of malware samples, including a wide range of families and variants. It encompasses both Windows and Android malware, offering a comprehensive view of the malware landscape.
- Advantages: The dataset is extensive and well-curated, facilitating the training of models on a broad spectrum of malicious behaviors.
- Challenges: The sheer volume of samples may pose challenges in terms of computational resources and processing time.

2. Microsoft Malware Classification Challenge (MS-Malware):
- Description: This dataset, released by Microsoft, contains a large-scale collection of labeled malware and benign samples. It was originally designed for a machine learning competition, providing a benchmark for evaluating detection models.
- Advantages: Well-labeled and balanced dataset suitable for training and evaluating supervised learning models.
- Challenges: As a snapshot in time, it may not fully capture the evolving nature of malware.

3. Contagio Mobile Malware Corpus:
- Description: Focused on mobile malware, this dataset includes samples targeting Android and other mobile platforms. It provides a valuable resource for developing and evaluating malware detection solutions for mobile devices.
- Advantages: Mobile-specific focus allows for targeted analysis and detection model training.
- Challenges: Limited to mobile malware, may not cover the full spectrum of threats affecting other platforms.

4. CICIDS2017 - Canadian Institute for Cybersecurity Intrusion Detection Systems 2017:
- Description: This network traffic dataset includes a variety of cyber threats, including malware, attacks, and normal traffic. It is designed for the evaluation of intrusion detection systems and can be adapted for malware detection.
- Advantages: Real-world network traffic data provides a holistic view of cyber threats.
- Challenges: Requires additional preprocessing for malware-specific analysis, as it is not tailored exclusively for malware detection.

5. UNSW-NB15:
- Description: The University of New South Wales Network-Based 15 (UNSW-NB15) dataset consists of network traffic data with a focus on intrusion detection. It includes both normal and malicious activities, making it suitable for evaluating network-based malware detection models.
- Advantages: Comprehensive network data with labeled instances of normal and malicious behavior.
- *Challenges:* Similar to CICIDS2017, may require additional preprocessing for malware-centric analysis.

6. Commonly Used Commercial Datasets:
- Description: Some commercial entities provide proprietary datasets tailored for training and evaluating their specific malware detection products. These datasets often reflect real-world threats faced by their users.
- Advantages: Reflects current and prevalent threats faced by users of specific security products.
- Challenges: Limited availability, potential bias towards threats prevalent in the targeted user demographic.

Selecting the appropriate dataset depends on the specific focus of the research or application, whether it be a general malware detection system, mobile-specific threats, or network-based intrusions. The subsequent sections of this paper will explore the metrics used to evaluate the performance of machine learning models, as well as the challenges and future directions in the field of malware detection.

Evaluation Metrics in Malware Detection Using Machine Learning:

The assessment of machine learning models for malware detection relies on various evaluation metrics that quantify their performance across different aspects. These metrics provide valuable insights into the model's

ability to correctly classify malware and benign instances, helping researchers and practitioners understand its strengths and weaknesses. Commonly used evaluation metrics include:

1. True Positive (TP):
   - *Definition:* The number of instances correctly identified as malware by the model.
   - *Interpretation:* A higher TP count indicates the model's effectiveness in correctly detecting actual malware.
2. True Negative (TN):
   - *Definition:* The number of instances correctly identified as benign by the model.
   - *Interpretation:* A higher TN count signifies the model's proficiency in correctly identifying non-malicious instances.
3. False Positive (FP):
   - *Definition:* The number of instances incorrectly identified as malware by the model when they are actually benign.
   - *Interpretation:* A higher FP count suggests a tendency of the model to produce false alarms, potentially leading to unnecessary interventions.
4. False Negative (FN):
   - *Definition:* The number of instances incorrectly identified as benign by the model when they are actually malware.
   - *Interpretation:* A higher FN count indicates instances where the model fails to detect actual malware, posing a risk of overlooking threats.
5. Precision:
   - *Calculation:* Precision = TP / (TP + FP)
   - *Interpretation:* Precision quantifies the accuracy of the model in correctly identifying malware. A higher precision indicates fewer false positives.
6. Recall (Sensitivity or True Positive Rate):
   - *Calculation:* Recall = TP / (TP + FN)
   - *Interpretation:* Recall measures the model's ability to capture all instances of actual malware. A higher recall indicates a lower risk of false negatives.
7. F1 Score:
   - *Calculation:* F1 Score = 2 * (Precision * Recall) / (Precision + Recall)
   - *Interpretation:* The F1 Score balances precision and recall, providing a comprehensive measure of the model's overall performance. It is particularly useful when there is an imbalance between malware and benign instances.
8. Specificity (True Negative Rate):
   - *Calculation:* Specificity = TN / (TN + FP)
   - *Interpretation:* Specificity assesses the model's capability to correctly identify non-malicious instances. A higher specificity indicates fewer false positives among benign instances.
9. Accuracy:
   - *Calculation:* Accuracy = (TP + TN) / (TP + TN + FP + FN)
   - *Interpretation:* Accuracy provides an overall measure of the model's correctness in classifying both malware and benign instances. However, it may be less informative in imbalanced datasets.
10. Area Under the ROC Curve (AUC-ROC):
    - *Interpretation:* The AUC-ROC metric evaluates the model's ability to distinguish between malware and benign instances across various threshold settings. A higher AUC-ROC score indicates better discriminatory power.

11. Area Under the Precision-Recall Curve (AUC-PR):
    - *Interpretation:* Similar to AUC-ROC, AUC-PR evaluates the precision-recall trade-off, especially in imbalanced datasets. A higher AUC-PR score signifies better precision-recall balance.

The selection of appropriate metrics depends on the specific goals and priorities of the malware detection task. Precision, recall, and F1 Score are particularly valuable in scenarios where imbalances exist between malware and benign instances. As machine learning models for malware detection continue to evolve, the careful consideration and interpretation of these metrics remain essential for assessing their efficacy.

Challenges and Limitations in Malware Detection Using Machine Learning:

Despite the promise and effectiveness of machine learning in malware detection, several challenges and limitations persist, influencing the overall efficacy and reliability of these systems.

1. Adversarial Attacks:

- *Challenge:* Malicious actors can deliberately manipulate features and patterns in malware to evade detection by machine learning models. Adversarial attacks aim to exploit vulnerabilities in the model's decision-making process.
- *Impact:* Adversarial attacks can significantly reduce the reliability of machine learning models, necessitating the development of robust and resilient detection mechanisms.

2. Imbalanced Datasets:
   - *Challenge:* Datasets used for training machine learning models often suffer from class imbalance, where the number of benign samples outweighs that of malicious samples or vice versa.
   - *Impact:* Imbalanced datasets can lead to biased models that favor the majority class, resulting in reduced sensitivity to the minority class (malicious samples) and an increased likelihood of false negatives.
3. Generalization Across Evolving Malware Variants:
   - *Challenge:* The rapid evolution of malware introduces new variants and behaviors, challenging machine learning models to generalize effectively across diverse and continuously changing threats.
   - *Impact:* Models trained on outdated or insufficiently diverse datasets may struggle to detect emerging malware variants, compromising the adaptability of the detection system.
4. Resource Constraints and Scalability:
   - *Challenge:* Machine learning models, especially deep learning models, can be computationally intensive, requiring substantial resources for training and inference.
   - *Impact:* Resource constraints can hinder the deployment and scalability of machine learning-based detection systems, particularly in environments with limited computational power.
5. Explainability and Interpretability:
   - *Challenge:* Machine learning models, particularly deep neural networks, are often perceived as "black boxes" due to their complex architectures. Understanding the decision-making process is challenging.
   - *Impact:* Lack of explainability and interpretability hinders the trustworthiness of the detection system, making it difficult to comprehend and justify the rationale behind specific predictions.
6. Feature Engineering and Selection:
   - *Challenge:* Identifying and extracting relevant features from dynamic and diverse malware samples is a non-trivial task. The effectiveness of the model is highly dependent on the quality of features used.
   - *Impact:* Inadequate feature engineering can lead to suboptimal model performance, emphasizing the need for ongoing research in feature extraction techniques.
7. Evolving Evasion Techniques:
   - *Challenge:* Malware creators continuously develop evasion techniques to bypass detection mechanisms, including those based on machine learning.
   - *Impact:* The arms race between detection systems and malware authors necessitates constant innovation in detection methods to counter evolving evasion techniques.

8. Privacy Concerns:
   - *Challenge:* Some machine learning approaches may involve the analysis of sensitive or personal data, raising privacy concerns.
   - *Impact:* Striking a balance between effective detection and protecting user privacy is essential to ensure the ethical use of machine learning in cybersecurity.

Understanding and addressing these challenges is crucial for advancing the capabilities of machine learning in malware detection. As the threat landscape continues to evolve, ongoing research and collaborative efforts are necessary to overcome these limitations and develop robust, adaptive, and privacy-conscious malware detection solutions.

Future Directions in Malware Detection Using Machine Learning:

The ever-evolving landscape of cybersecurity and the persistent threat of malware demand continuous innovation and adaptation in detection methodologies. Future directions in the field of malware detection using machine learning are poised to address existing challenges and leverage emerging technologies for enhanced effectiveness. Several key areas represent promising avenues for future research and development:

1. Explainable AI and Interpretability:
   - *Direction:* Future efforts should prioritize the development of machine learning models with enhanced explainability and interpretability. Understanding the decision-making process is critical for building trust in detection systems, especially in sensitive environments.
2. Ensemble Learning Approaches:

- *Direction:* Ensemble learning, which combines predictions from multiple models, offers a robust approach to enhance detection accuracy and resilience against adversarial attacks. Future research should explore and optimize ensemble techniques for malware detection.

3. Real-time and Proactive Detection:
   - *Direction:* The transition towards real-time and proactive detection mechanisms is essential to minimize the impact of rapidly spreading malware. Future systems should aim for low-latency detection and adaptive responses to emerging threats.
4. Integration with Threat Intelligence Feeds:
   - *Direction:* Incorporating threat intelligence feeds into machine learning models enhances their awareness of the current threat landscape. Future systems should seamlessly integrate with external threat intelligence sources for more context-aware and accurate detection.
5. Deep Learning Architectures:
   - *Direction:* Advancements in deep learning architectures, including novel neural network structures and training techniques, hold promise for improving the ability to capture complex patterns in malware behavior. Future research should explore architectures optimized for both accuracy and efficiency.
6. Transfer Learning and Few-shot Learning:
   - *Direction:* Transfer learning, leveraging knowledge gained from one domain to improve learning in another, and few-shot learning, training models with minimal examples, are emerging areas. Future efforts should investigate their application to malware detection, especially in scenarios with limited labeled data.
7. Examination of Network Traffic:
   - *Direction:* With the increasing sophistication of malware using network-based attacks, future research should focus on advanced techniques for analyzing network traffic. This includes the development of models capable of identifying anomalous patterns indicative of malware communication.
8. Human-in-the-loop Approaches:
   - *Direction:* Integrating human expertise into the detection process through human-in-the-loop approaches can enhance the adaptability of systems. Future systems may leverage human feedback to improve model performance and address challenges such as explainability.
9. Quantum Computing for Cryptographic Analysis:
   - *Direction:* As quantum computing advances, future research should explore its application to cryptographic analysis, enabling more effective detection of malware leveraging advanced cryptographic techniques.
10. Ethical and Privacy-conscious Designs:
    - *Direction:* Future developments must prioritize ethical considerations and privacy-conscious designs. Striking a balance between effective detection and protecting user privacy is essential for the responsible deployment of machine learning in cybersecurity.

Continued collaboration between researchers, industry experts, and cybersecurity professionals is paramount for the success of these future directions. As the threat landscape evolves, embracing innovative approaches and technologies will be crucial in staying ahead of sophisticated and adaptive malware.

Conclusion:

In the ever-evolving landscape of cybersecurity, the battle against malware remains a dynamic and complex challenge. This paper has provided a comprehensive exploration of the integration of machine learning techniques into the realm of malware detection. The discussion has spanned traditional detection methods, the application of various machine learning algorithms, feature extraction techniques, datasets for training and evaluation, and the challenges and future directions in this field.

Traditional methods, such as signature-based and heuristic-based detection, have laid the groundwork for cybersecurity but face limitations in adapting to the rapid evolution of malware. Machine learning, with its capacity to learn from data and identify intricate patterns, offers a promising paradigm shift. Supervised learning, unsupervised learning, and deep learning algorithms, when coupled with effective feature extraction, showcase the potential to enhance detection accuracy and adaptability.

Feature extraction, encompassing static, dynamic, and hybrid analysis, serves as a critical bridge between raw data and machine learning models. The quality and relevance of extracted features influence the efficacy of detection systems, making ongoing research in this area pivotal for advancements in malware detection.

Datasets, ranging from the Malware Genome Project to commercial datasets, provide the foundation for training and evaluating machine learning models. However, challenges such as class imbalance and the need for representative datasets underscore the importance of careful dataset selection and curation.

While machine learning holds great promise, challenges persist, including adversarial attacks, imbalanced datasets, and the need for explainability. Addressing these challenges and moving towards real-time, proactive,

and privacy-conscious detection mechanisms represents the future of malware detection. Ensemble learning, integration with threat intelligence, and advancements in deep learning architectures are among the key directions for future research.

As the cybersecurity landscape continues to evolve, the collaboration of researchers, practitioners, and industry experts is essential for staying ahead of sophisticated and adaptive malware. Future systems must not only focus on detection accuracy but also emphasize ethical considerations, privacy protection, and the interpretability of models. The ongoing pursuit of innovative solutions and the integration of cutting-edge technologies will play a pivotal role in fortifying our defenses against the ever-present threat of malware.

References:

1. D. Marcozzi, M. Colajanni, A. C. Mancini, and C. Pazzaglia, "Machine learning techniques for the detection of malicious executables," in *Computers & Security*, vol. 30, no. 3, pp. 241-250, May 2011.

2. W. Hu, K. Zhang, R. Huang, and C. K. Hui, "Malware detection through machine learning using dynamic analysis features," in *Computers & Security*, vol. 59, pp. 226-238, May 2016.

3. I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *IEEE Communications Surveys & Tutorials*, vol. 17, no. 3, pp. 1978-2000, 2015.

4. A. Kolosnjaji, A. Zarras, C. Demontis, E. Biggio, G. Giacinto, I. Corona, and F. Roli, "Adversarial malware binaries: Evading deep learning for malware detection in executables," in *arXiv preprint arXiv:1705.07101*, 2017.

5. H. Han, W. Fan, and P. S. Yu, "Mining useful knowledge from unstructured text data," in *Advances in Knowledge Discovery and Data Mining*, pp. 515-520, Springer, 2001.

6. M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proceedings of the Second IEEE International Conference on Computational Intelligence for Security and Defense Applications 2009*, pp. 53-58, IEEE, 2009.

7. K. Grosse, N. Papernot, P. Manoharan, M. Backes, and P. McDaniel, "Adversarial perturbations against deep neural networks for malware classification," in *arXiv preprint arXiv:1606.04435*, 2016.

8. S. Saxe and J. Berlin, "eXpose: A character-level convolutional neural network with highway layers for malware detection," in *arXiv preprint arXiv:1606.04435*, 2015.

9. Y. Zhang, W. Cui, S. Member, J. Wang, and R. H. Deng, "Malware detection based on dynamic features of API calls," in *International Conference on Information and Communications Security*, pp. 441-452, Springer, 2015.

10. T. M. Mitchell, "Machine learning," McGraw Hill, vol. 45, no. 4, p. 9, 1997.