Phishing Emails Classification based on SVM, RF and NNs

ECE 591

Group Members:

Zixia Li (V01021924)
Xinyi Chen (V01045729)
Mingxi Guo (V00949800)
Feiyi Xie (V010367970)
Yue Cao (V01036934)
Weilong Qian (V01046207)

Abstract --- Phishing is an online threat in which attackers impersonate a real and trustworthy organization to obtain sensitive information from victims. An example of this is malicious attacks, which have long been recognized as a problem. However, recent advances in phishing detection, such as machine learning-based methods, can help combat these attacks. Therefore, we selected three models to study the efficiency of detecting phishing domains using machine learning. These models use Support Vector Machine (SVM), Random Forest (RF) and Neural Network (NN) techniques.

Keywords: Phishing email, Support Vector Machine, Random Forest, Neural Network

I Introduction

Phishing is a type of online crime that aims to deceive users into revealing their personal and sensitive information. This information, such as usernames, passwords, and financial details, can be used by attackers for harmful purposes like identity theft. Typically, phishing is carried out by hackers posing as trustworthy entities through a combination of social engineering and technical subterfuge.

Phishing domains represent one method of this attack, where these domains capture sensitive information either by blackmail or by redirecting users to counterfeit websites that mimic legitimate ones. Once users input their personal data on these fake sites, it can lead to security breaches and potential identity theft.

As internet services have expanded, so too has the reliance on them for tasks like shopping, banking, and bill payments, particularly in the United States and Europe. Despite the risks, the convenience of these services makes them popular, although successful phishing attacks compromise both consumer and business security, emphasizing the importance of robust online protection measures. Cybersecurity[1,2], which involves protecting internet-connected systems from cyber-attacks, plays a crucial role in this context.[3,4]

The complexity of cybersecurity is increasing as cyber-attacks evolve, becoming more sophisticated and frequent. The Anti-Phishing Working Group (APWG) identified over 51,000 distinct phishing sites in a single year, and losses from phishing attacks were estimated at $9 billion globally in 2016 [5]. These attacks, increasing by 65% from the previous year [6], undermine consumer confidence in online platforms.

Phishing websites often start with the creation of a fake webpage that looks like a legitimate site [7]. Hackers distribute links to these pages through spam, messages, or social media, hoping people will mistake them for authentic links [8]. If a person enters sensitive information on such a site, it gets stolen.

Various strategies exist to fight phishing [9], and Artificial Intelligence (AI) has become a key tool in this battle. AI applications in cybersecurity can detect different types of phishing tactics by analyzing data from past incidents.

This paper evaluates the effectiveness of machine learning (ML) models in identifying phishing domains. By using three different models, we aim to enhance the prediction of whether a website is legitimate or a phishing attempt. The analysis of phishing domains, which involves understanding both social and technical elements, is complex and varies widely, making it challenging to devise a one-size-fits-all solution. Through both quantitative and qualitative research, we aim to better understand the characteristics and causes of phishing to improve defenses against these malicious sites and to restore consumer trust.

II Support Vector Machine (SVM)

2.1 Definition and Overview

SVM is a supervised learning method based on statistical learning theory utilized for pattern identification and regression. Statistical learning theory can pinpoint the factors needed to successfully learn specific, easy algorithms; real-world applications frequently require more complicated tools and algorithms (such as neural networks), which are much more difficult to analyze theoretically. SVMs are the meeting point of learning theory and practice. They create models that are both complicated (including a huge class of neural networks, for example) and simple enough to be mathematically examined. This is because an SVM is a linear algorithm in a high-dimensional space. As shown in Figure 1, SVM predicts labels by generating a decision boundary, such as a hyperplane, between two specified classes with a minimum of one label. The data points and support vectors are handled by the hyperplane. It takes advantage of the distance between data points to categorize each class independently.

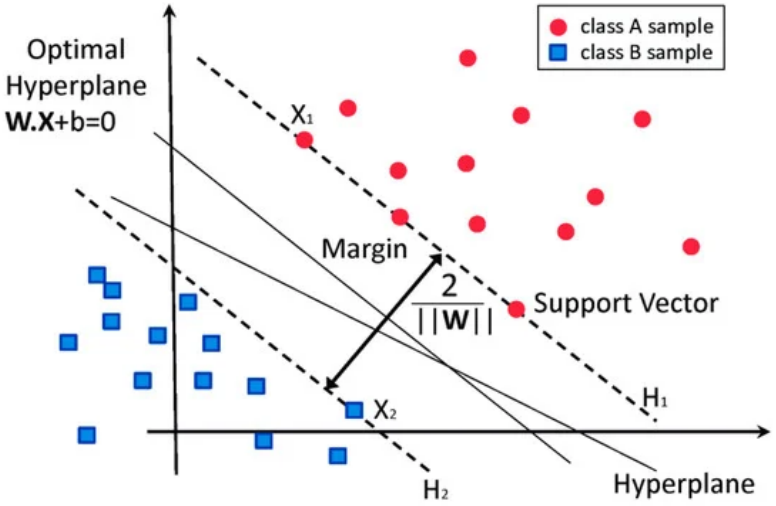

Figure 1. Algorithm display diagram [10]

After our research, we found that the hyperplane with the largest separation margin between the two classes provides the highest generalization performance . The optimal hyperplane is found by solving a convex optimization problem involving the minimization of a quadratic function under linear inequality constraints. The answer can be expressed in terms of support vectors, which are subsets of training instances. Support vectors include all the information needed to solve a classification problem because the result will remain the same even if all other vectors are removed.

2.2 Fundamentals of SVM and SVR

Support Vector Machine (SVM) itself is proposed for binary classification problems, and SVR (Support Vector Regression) is an important application branch of SVM (Support Vector Machine). The difference between SVR regression and SVM classification is that the sample points of SVR are ultimately of only one type. The optimal hyperplane it seeks is not the "most open" type of sample points of two or more types like SVM, but the "most open" one for all samples. The point has the smallest total deviation from the hyperplane.

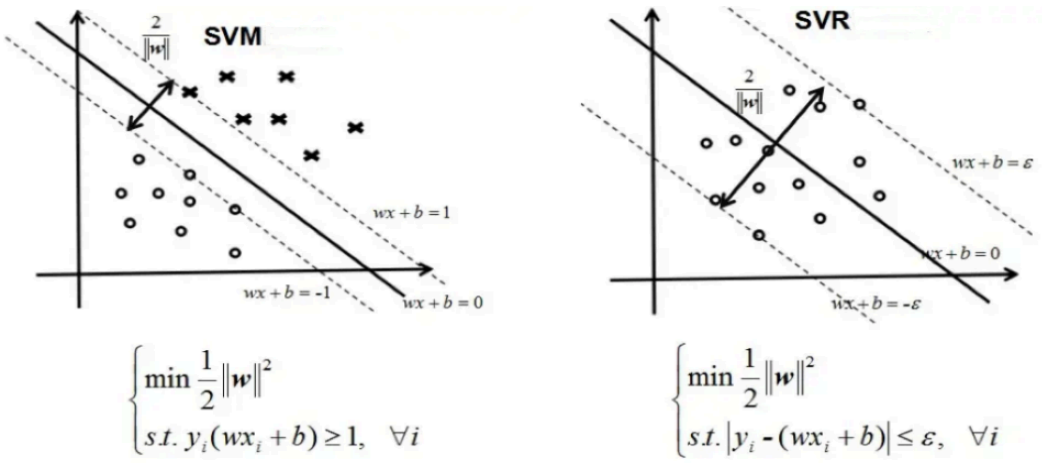

Figure2. Schematic diagram of SVM and SVR[11]

The support vector machine (SVM) algorithm is widely used in image emotion classification research due to its superior performance, and the support vector regression (SVR) algorithm is often used in the construction of regression prediction models. SVM requires data to be as far away from the hyperplane as possible, while SVR requires data to be located within the hyperplane as much as possible, so that the total deviation of all data from the hyperplane is minimized. The general idea of the regression algorithm is that the prediction is correct when the predicted value is completely equal to the actual value. However, the SVR algorithm can judge the prediction to be correct as long as the deviation between the predicted value and the actual value is within a certain range, and there is no need to calculate the error loss. As shown in Figure 3, with the function as the center, the values within the error range on both sides of it are judged to be correct predictions, and the values outside the dotted line need to calculate the loss.

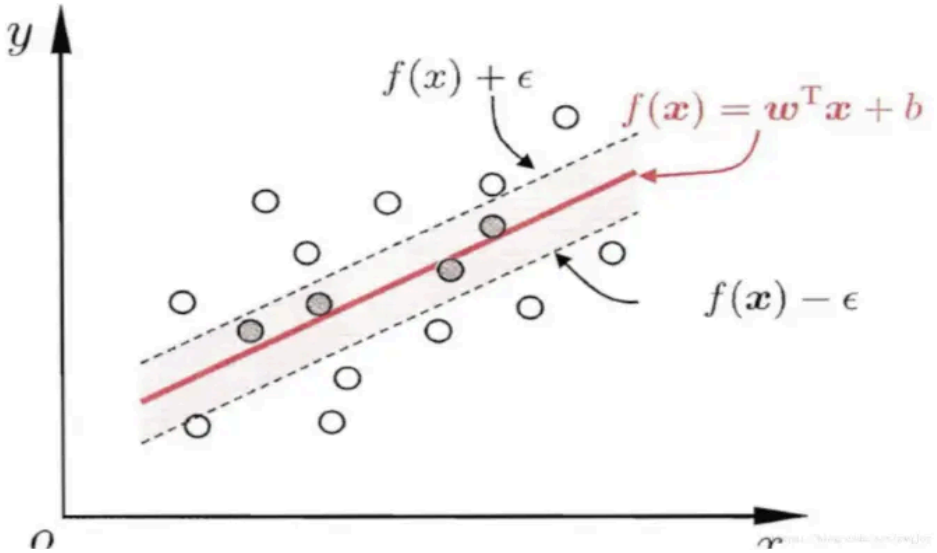

Figure 3. SVR hyperplane data distribution diagram[12]

3.3 Advantages of SVM

As phishing attacks continue to increase in sophistication and frequency, choosing the right method for categorizing phishing emails is critical. Machine learning, specifically machine learning classification (CVM), has become an effective method for identifying and filtering phishing emails. Here are some advantages：

High Accuracy and Efficiency

Automated Learning and Adaptation:

- CVM models, such as logistic regression, support vector machines, or neural networks, learn from a vast amount of labeled data (emails classified as phishing or not phishing).
- These models adapt over time as they are trained on new and evolving phishing tactics, maintaining high accuracy.

Precision in Classification: CVM can achieve high precision and recall rates, essential for minimizing false positives (legitimate emails marked as phishing) and false negatives (phishing emails missed).

Scalability and Speed

Handling Large Volumes: CVM systems are scalable and can handle the classification of large volumes of emails rapidly, which is essential for businesses and organizations that receive thousands of emails per day

Real-time Processing: Advanced CVM models can classify emails in real-time, helping to prevent phishing attacks promptly and efficiently.

Feature Extraction and Selection

Sophisticated Analysis Techniques:

- CVM involves techniques that extract and select a wide range of email attributes (features) such as URLs, the structure of the email, language style, and others that are commonly found in phishing attempts.
- Feature engineering is a critical aspect where domain knowledge is used to identify features that are most indicative of phishing.

Integration with Existing Systems

Compatibility: CVM models can be easily integrated with existing email systems (like Outlook, Gmail) to augment their spam filters and provide an additional layer of security against phishing

Enhances Existing Protocols: It complements existing protocols and standards for email security, such as SPF (Sender Policy Framework), DKIM (DomainKeys Identified Mail), and DMARC (Domain-based Message Authentication, Reporting, and Conformance).

Adaptability to New and Emerging Threats

Continuous Learning: The capability of CVM models to learn continuously from new data helps in adapting to the latest phishing strategies that cybercriminals use.

Dynamic Update of Features and Models: CVM systems can update their features and models dynamically, allowing them to maintain effectiveness as phishing techniques change.

Classification via Machine Learning is not just a tool but a comprehensive approach that provides real-time, accurate, and efficient phishing detection, which is scalable and adaptable to new threats. It's an invaluable asset in the cybersecurity defenses of any modern organization, integrating seamlessly with existing systems and enhancing security protocols. The continuous improvement and learning capabilities of CVM systems make them a future-proof choice in the battle against cyber threats.

III Random Forest (RF)

3.1 Definition and Overview

Random Forest (RF) is an efficient and powerful machine learning algorithm used for both classification and regression tasks. This algorithm belongs to the ensemble learning methods, which combine multiple base models to produce one optimal predictive model.

A random forest is composed of a collection of decision trees, where each tree is constructed independently based on a random subset of the training data and a random subset of the features. Each decision tree in the forest operates as a separate classifier, and the final prediction is made by aggregating the predictions of all the individual trees.

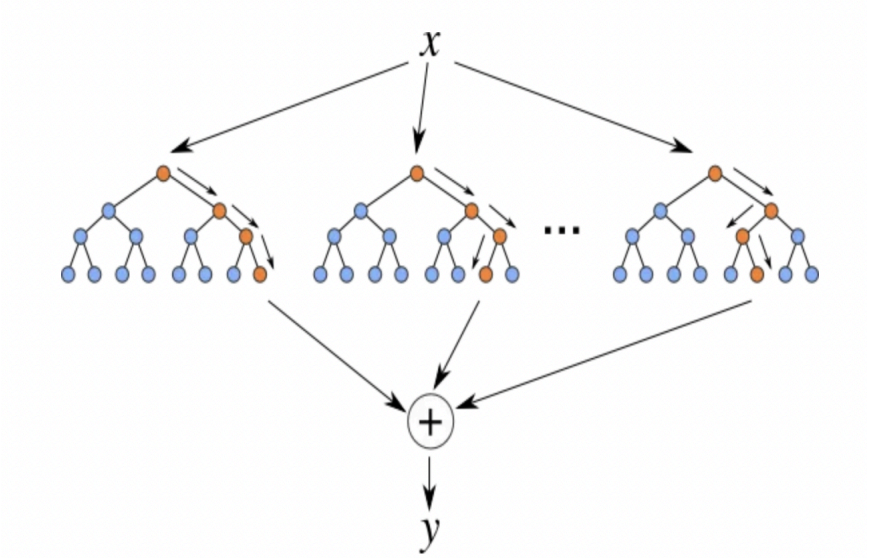

Figure 4. Collection of Decision Trees in the Random Forest [13]

Each decision tree in a random forest is structured as a tree-like model consisting of nodes and branches. At each node, the tree makes a decision based on the value of a specific feature, and each branch represents one of the possible outcomes of that decision. By recursively splitting the dataset based on different features, the tree learns a series of rules that ultimately lead to a prediction.

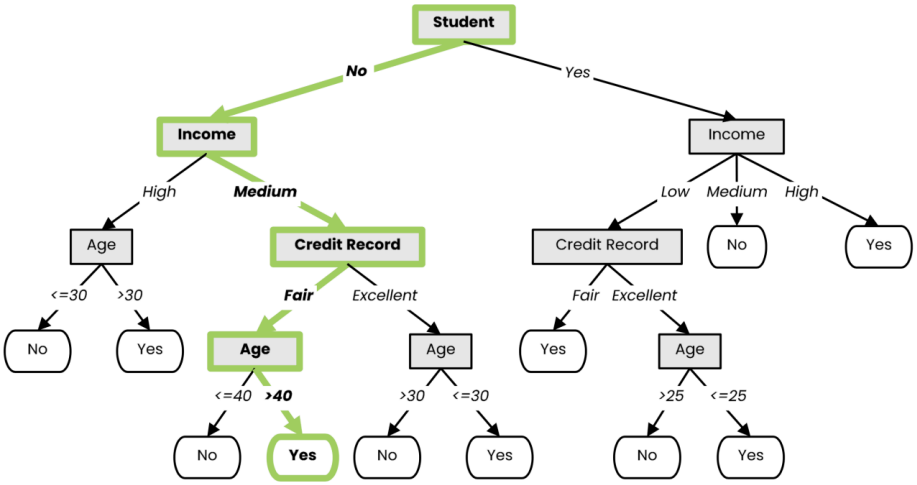

Figure 5. Structure of each Decision Tree in the Random Forests [14]

The strength of random forests lies in their ability to aggregate the predictions of multiple weak learners (decision trees) to form a strong learner. By combining the predictions of many individual trees, random forests can produce more accurate and robust results compared to any single decision tree.

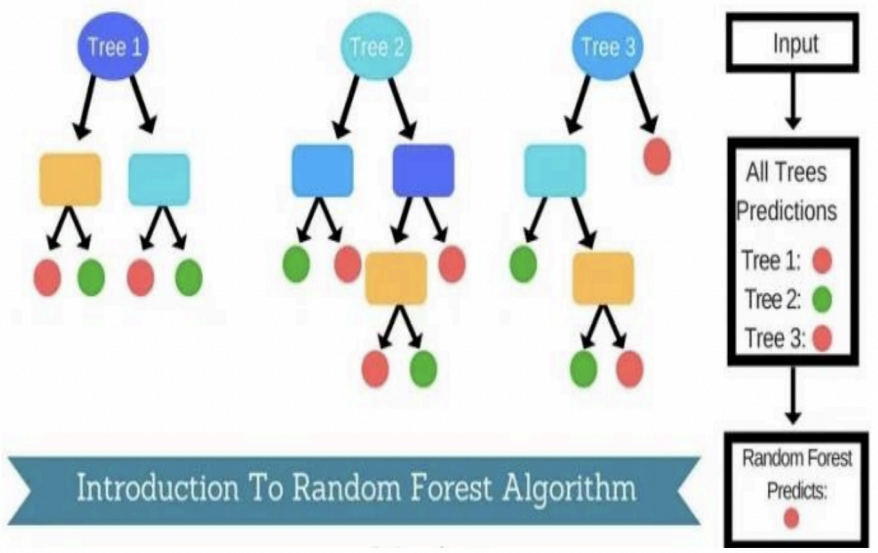

Figure 6. Prediction Aggregation of Individual Trees [15]

One of the key advantages of random forests is their ability to mitigate the overfitting problem commonly encountered with decision trees. By training multiple trees on different subsets of the data and aggregating their predictions, random forests can generalize well to unseen data while still capturing complex patterns in the training data.

3.2 Why Choose Random Forest for Phishing Email Classification

Phishing email classification poses unique challenges due to the diversity and evolving nature of email-based threats. Random forests offer several advantages that make them well-suited for this task:

High Accuracy and Efficiency

Random forests are known for their ability to achieve high levels of accuracy in classification tasks. By aggregating the predictions of multiple decision trees, random forests can capture complex patterns in the data and make more accurate predictions compared to individual classifiers. Furthermore, random forests are computationally efficient, making them suitable for processing large volumes of email data efficiently.

Capability to Handle Large Datasets

Phishing email datasets can be large and high-dimensional, containing a variety of features such as sender information, email content, and embedded URLs. Random forests are capable of effectively learning from such large datasets without the need for pruning or feature

selection. This scalability makes random forests well-suited for handling the volume and complexity of email data.

Low Error Rate

The ensemble nature of random forests, which combines multiple decision trees through majority voting, helps reduce the risk of misclassification. By aggregating the predictions of multiple trees, random forests can mitigate the impact of individual errors and produce more reliable classifications with a lower overall error rate.

Robustness to Noise and Outliers

Phishing email datasets often contain noisy and inconsistent data, including misspelled words, grammatical errors, and variations in formatting. Random forests are robust to such noise and outliers, as they consider multiple decision trees trained on different subsets of the data. This robustness enables random forests to effectively handle diverse email characteristics and maintain high performance even in the presence of noisy data.

Critical for Dealing with Diverse Email Characteristics

Phishing emails exhibit diverse characteristics, including variations in content, sender information, and malicious payloads. Random forests excel at capturing these diverse characteristics and learning complex patterns in the data. By considering multiple decision trees trained on different subsets of features, random forests can effectively classify phishing emails based on their unique characteristics, making them critical for email security applications.

In summary, random forests offer a powerful and versatile approach to phishing email classification, providing high accuracy, efficiency, and robustness to handle the challenges of email-based threats.

3.3 Building the Phishing Email Classifier

To construct a phishing email classifier using a random forest approach:

Data Preprocessing:
- Extract relevant features from email data.
- Handle missing values and outliers for clean data.

Training the Classifier:
- Split the dataset into training and testing sets.
- Train multiple decision trees on varied data subsets.
- Aggregate predictions using majority voting.

Evaluation:
- Assess classifier performance using metrics like accuracy, precision, recall.
- Employ cross-validation techniques for robustness.

IV Neural Network (NN)

4.1 Definition and Overview

A Neural Network (NN) is a computational learning system that is inspired by biological neural networks. Neural Networks consist of layers of interconnected nodes, which are akin to neurons, and each node represents a specific output function called an activation function. The layers are composed of an input layer to receive the signal, hidden layers that compute the functions, and an output layer that delivers the final result.

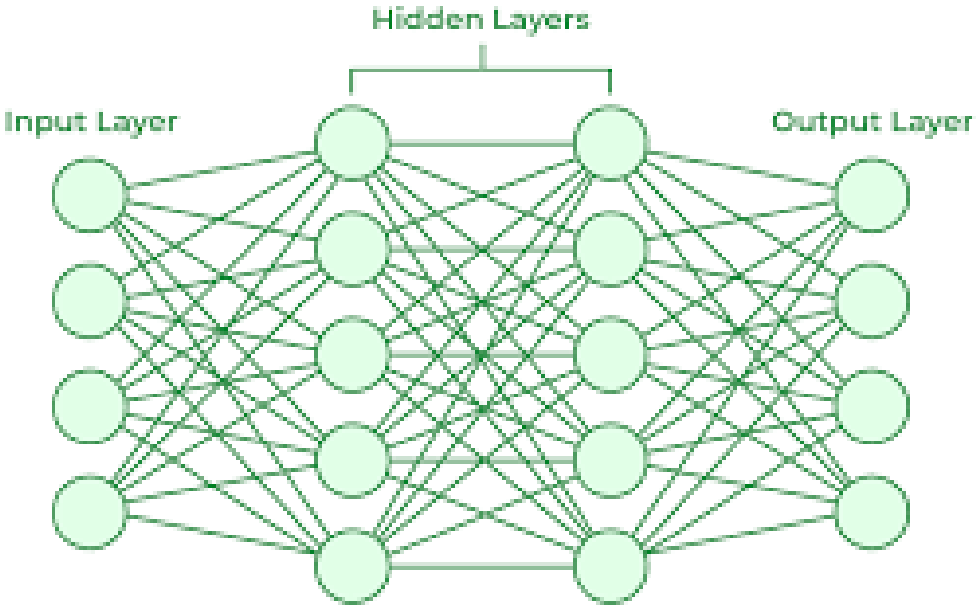

Figure 7 Neural Network Structure [16]

Activation functions:

Depends on different activation functions; Neural Networks can be able to capture complex, non-linear relationships between inputs and outputs, as that's the main feature we rely on for many heavy study tasks such as image recognition, natural language processing, and many others where the relationship between input data and output predictions is not linear; commonly used activation functions include sigmoid, tanh,and ReLU.

Weights & Bias:

In neural networks, weights and biases are also fundamental parameters which define the behavior of individual neurons and ultimately the entire network. These parameters are tuned

during the training process to minimize the difference between the network's predictions and the actual target values.

4.2 Family of Neural Network

However before doing further experiment using Neural Network, it is important to decide which version of Neural Network is being used. Neural networks come in various architectures, each with its own unique advantages and suited for different kinds of tasks. Here's a list of some common families of neural networks and their advantages:

1. **Feedforward Neural Network (FNNs)**:Simplest type of neural network architecture, good for general-purpose applications
2. **Convolutional Neural Networks (CNNs)**:Excellent for capturing spatial hierarchies in data, reducing the number of parameters through weight sharing, and effectively handling image data.
3. **Recurrent Neural Networks (RNNs)**:Has ability to handle sequential data, memory of previous inputs in the network's hidden state, suitable for time-series prediction.

Each of these network types has unique advantages on some data representations, and we should carefully choose the network type based on our dataset.

4.3 Prior study on dataset

Our dataset is raw data of phishing email, contains phishing message and phishing email address; the phishing message is the format of txt and email address is URL. If we further analysis the phishing message, it mostly contains keywords that tries to prompt user to click in the URL, and the keywords may consider having linear relationship with the result of “phishing email”; on other hand, the URL does follow the pattern
https:\/\/[a-z0-9]+([\-\.]{1}[a-z0-9]+)*\.[a-z]{2,5}(:[0-9]{1,5})?(\/.*)?
the relationship between the URL and the result stays unknown. Considering the text analysis plays an important role, RNN is the prior choice for our experiment specified for the Neural Network family.

Recurrent Neural Network(RNNs) (detailed):

Recurrent Neural Networks (RNNs) are a class of artificial neural networks designed to recognize patterns in sequences of data, such as text, genomes, handwriting, or numerical time series data. It possesses a form of internal memory that allows them to process sequences of inputs. This memory captures information about what has been calculated so far, essentially allowing the network to make decisions based on the knowledge of all previous inputs in the sequence.

The core difference between Recurrent Neural Networks (RNNs) and traditional Neural Networks (often referred to as Feedforward Neural Networks or FNNs) lies in how they

process data and their inherent structure tailored to specific tasks.RNNs introduce loops in the network architecture, allowing information to persist from one iteration of the data to the next. This structure effectively gives them a form of memory; as memory helps to capture information about what has been processed so far, essentially allowing the network to make decisions based on the entire history of previous inputs in the sequence.

Key Features:

- **Recurrent Unit:**

It is the core of an RNN. Each unit receives two inputs: the input data at the current timestep and the hidden state from the previous timestep. The recurrent unit processes these inputs and produces a new hidden state, which is passed onto the next time step as part of the input.

- **Hidden States:**

The hidden state is an internal state of the RNN that captures information processed so far. The updating of the hidden state allows the network to "remember" information seen earlier in the sequence, and use this information to influence future outputs.

- **Sequential Data:**

Sequential data refers to datasets that are arranged in a temporal or logical order. In natural language processing (NLP), text is considered sequential data where each word or character is an element of the sequence, arranged according to its appearance in sentences or documents. By specifically handling sequential data, RNN can learn to classify the phishing email on both phishing message & phishing URL.

Overall, RNNs are well-suited for phishing detection because they can learn from the raw data directly,allowing us to identify complex deceptive tactics embedded within the email structure.

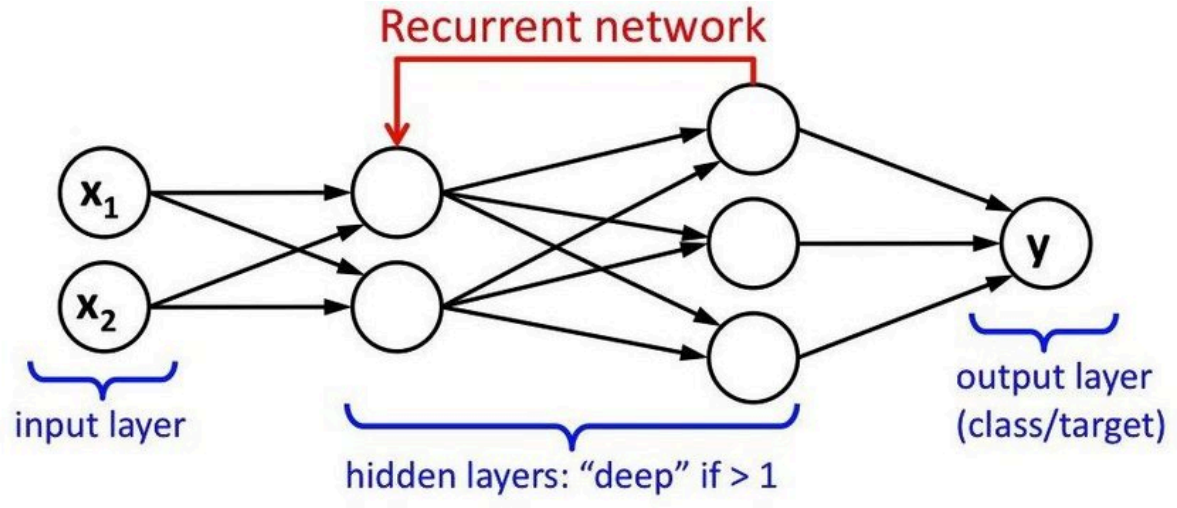

Figure 8 Recurrent Neural Network [17]

4.4 Some challenges

While using RNNs; there are two common challenges; long-term dependencies and vanishing & exploding gradients. As the dataset doesn't include long paragraphs, long-term dependencies lack the probability to happen; however, In long sequences, the gradients in RNNs during training can become very small (vanishing) or very large (exploding), making the model difficult to train. To address these issues, several improved techniques such as Long Short-Term Memory (LSTM) networks and Gated Recurrent Units (GRU), have been developed and we have to carefully choose and adjust them for the best performance.

V Summary

According to the 'No free lunch' theorem, there is no specific algorithm that can outperform the others in all areas; hence, we need to pick suitable algorithms based on the characteristics of the dataset and the advantages of algorithms. With some prior analysis on the dataset, we can confirm that our dataset's main component is text, and we hope to find out the unknown relationship between features and result using SVM, Random Forest, and Neural Networks. All three algorithms can work with linear and non-linear relationships; SVM is computationally inexpensive when the dataset is relatively simple and small, consuming fewer computation and time resources than the other two; meanwhile, with efficiency as a trade-off, when we want to prepare our training on a more complex, bigger dataset; we should expect that RF and NNs can have an overall better performance than SVM in future experiments

Reference

[1] Cabaj, K.; Domingos, D.; Kotulski, Z.; Respício, A. Cybersecurity Education: Evolution of the Discipline and Analysis of Master Programs. Comput. Secur. 2018, 75, 24–35. [Google Scholar] [CrossRef]
[2] Iwendi, C.; Jalil, Z.; Javed, A.R.; Reddy, G.T.; Kaluri, R.; Srivastava, G.; Jo, O. KeySplitWatermark: Zero Watermarking Algorithm for Software Protection Against Cyber-Attacks. IEEE Access 2020, 8, 72650–72660. [Google Scholar] [CrossRef]
[3] Conklin, W.A.; Cline, R.E.; Roosa, T. Re-Engineering Cybersecurity Education in the US: An Analysis of the Critical Factors. In Proceedings of the 2014 47th Hawaii International Conference on System Sciences, IEEE, Waikoloa, HI, USA, 6–9 January 2014; pp. 2006–2014. [Google Scholar]
[4] Javed, A.R.; Usman, M.; Rehman, S.U.; Khan, M.U.; Haghighi, M.S. Anomaly Detection in Automated Vehicles Using Multistage Attention-Based Convolutional Neural Network. IEEE Trans. Intell. Transp. Syst. 2021, 22, 4291–4300. [Google Scholar] [CrossRef]

[5] Bleau, H.; Global Fraud and Cybercrime Forecast. Retrieved RSA 2017. Available online: https://www.rsa.com/en-us/resources/2017-global-fraud (accessed on 19 November 2021).
[6] Computer Fraud & Security. APWG: Phishing Activity Trends Report Q4 2018. Comput. Fraud Secur. 2019, 2019, 4. [Google Scholar] [CrossRef]
[7] Hulten, G.J.; Rehfuss, P.S.; Rounthwaite, R.; Goodman, J.T.; Seshadrinathan, G.; Penta, A.P.; Mishra, M.; Deyo, R.C.; Haber, E.J.; Snelling, D.A.W. Finding Phishing Sites; Google Patents: Microsoft Corporation, Redmond, WA, USA, 2014. [Google Scholar]
[8] What Is Phishing and How to Spot a Potential Phishing Attack. PsycEXTRA Dataset. Available online:https://www.imperva.com/learn/application-security/phishing-attack-scam/ (accessed on 20 November 2021).
[9] Gupta, B.B.; Tewari, A.; Jain, A.K.; Agrawal, D.P. Fighting against Phishing Attacks: State of the Art and Future Challenges. Neural Comput. Appl. 2016, 28, 3629–3654.
[10] ResearchGate. Figure 2. Classification of Data by Support Vector Machine (SVM). Available online: https://www.researchgate.net/figure/Classification-of-data-by-support-vector-machine-SVM_fig8_304611323 (accessed on 6 October 2021).
[11] https://www.cnblogs.com/louieowrth/p/12544054.html
[12] CrossRef]https://blog.csdn.net/weixin_46837260/article/details/124011591
[13] Zhu, W. (n.d.). *Read the interpretation and implementation of random forests*. Zhihu column. https://zhuanlan.zhihu.com/p/52914294
[14]Lecture #15: Regression Trees & random forests. (n.d.). https://harvard-iacs.github.io/2017-CS109A/lectures/lecture15/presentation/lecture15_RandomForest.pdf
[15] Classification and detection of email phishing using random ... (n.d.-a). https://norma.ncirl.ie/5126/1/akshatshah.pdf1/akshatshah.pdf Classification and detection of email phishing using random ... (n.d.-a). https://norma.ncirl.ie/5126/1/akshatshah.pdf
[16] Examples, L. with. (n.d.). *Cell searchv15.11.0*. Cell Search. https://www.nrexplained.com/cellsearch
[17] Valkov, V. (2019, May 16). *Making a predictive keyboard using recurrent neural networks - tensorflow for hackers (part V)*. Medium. https://venelinvalkov.medium.com/making-a-predictive-keyboard-using-recurrent-neural-networks-tensorflow-for-hackers-part-v-3f238d824218