# Phishing Emails Classifiers Research

- Group 8

University of Victoria

# Introduction

**Phishing email attack:**

a fraudulent attempt to trick an email recipient into sharing sensitive information.

The sender poses as a reputable business or known person in order to get the recipient to click on a link and open an attachment. [1]

**Our goal:** find out good classifiers to classify phishing emails

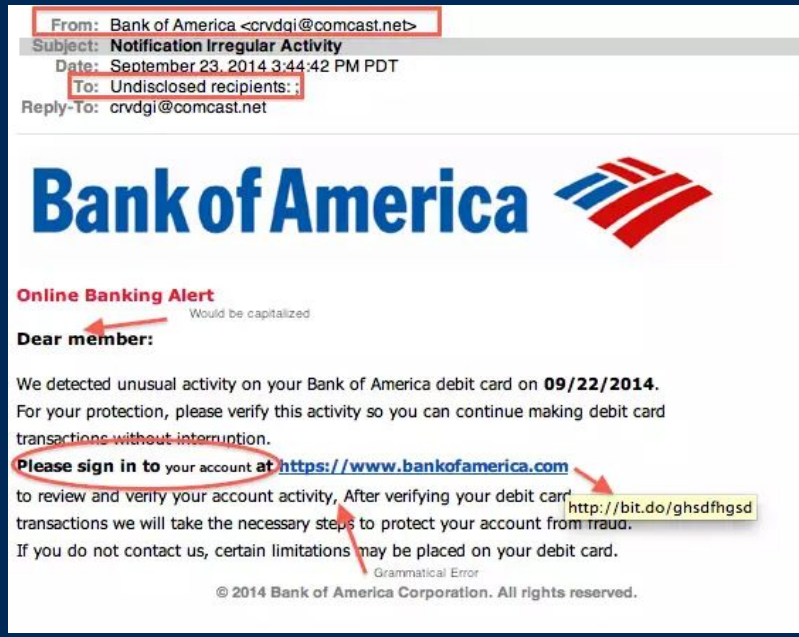



Fig 1. An phishing email example [1]

# Random Forest - Introduction

- Definition [2]
  - Random forests (random decision forests)
  - For classification, regression, and other tasks
  - Output:
    - Classification - the classes/categories
    - Regression - mean prediction
  - An ensemble learning method
    - Is made up of a set of classifiers, e.g. decision trees
    - Predictions are aggregated to identify the most popular result

# Random Forest - Individual Decision Tree

- A tree-like model that illustrates series of events leading to certain decisions
- Each node represents a test on an attribute and each branch is an outcome of that test

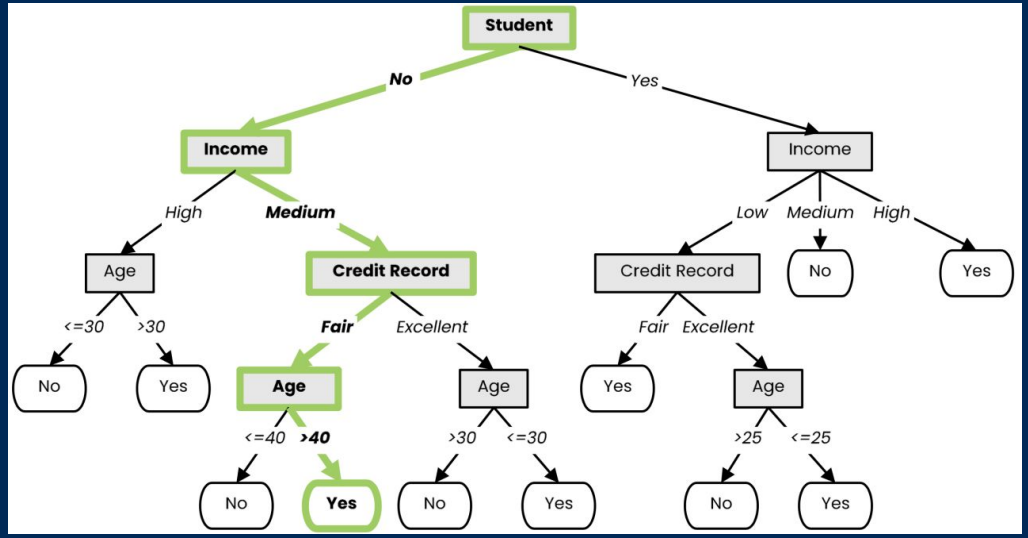
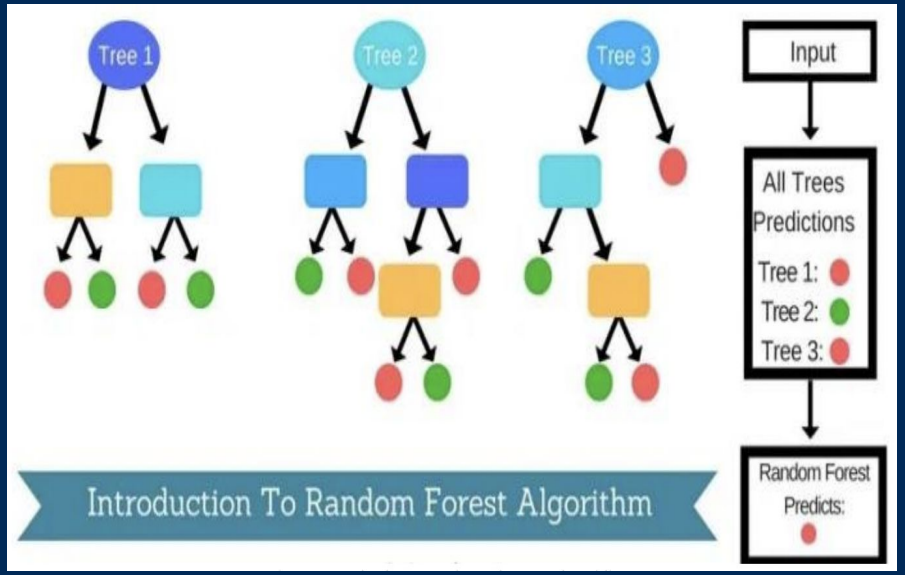



Fig 2. sample for Decision tree classifier [3]

# Random Forest Structure

- After a large number of trees is generated, they vote for the most popular class. We call these procedures random forests
- Address the problem of decision tree overfitting problem



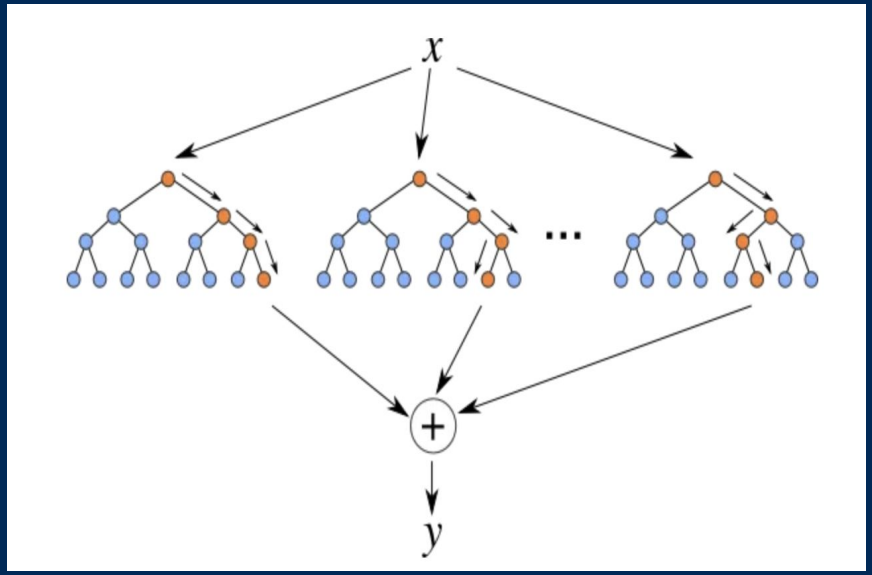




Fig 3. sample for Random forest classifier [4]

# Why Choose Random Forest for Phishing Email Classification?

- High accuracy and efficiency [5, 6]
  - Capable of effective learning in large datasets, handling high-dimensional data without pruning, minimizes overfitting
- Low error rate
  - Utilize multiple decision trees
  - Majority voting reduces risk of misclassification
- Robustness to noise and outliers
  - Naturally handles noisy and inconsistent data
  - Critical for dealing with diverse email characteristics

Cons: less interpretable, computational complexity [7]

# Neural Network

Layers

- Input Layer
- Hidden Layers
- Output Layer

Weights and biases

Activation functions

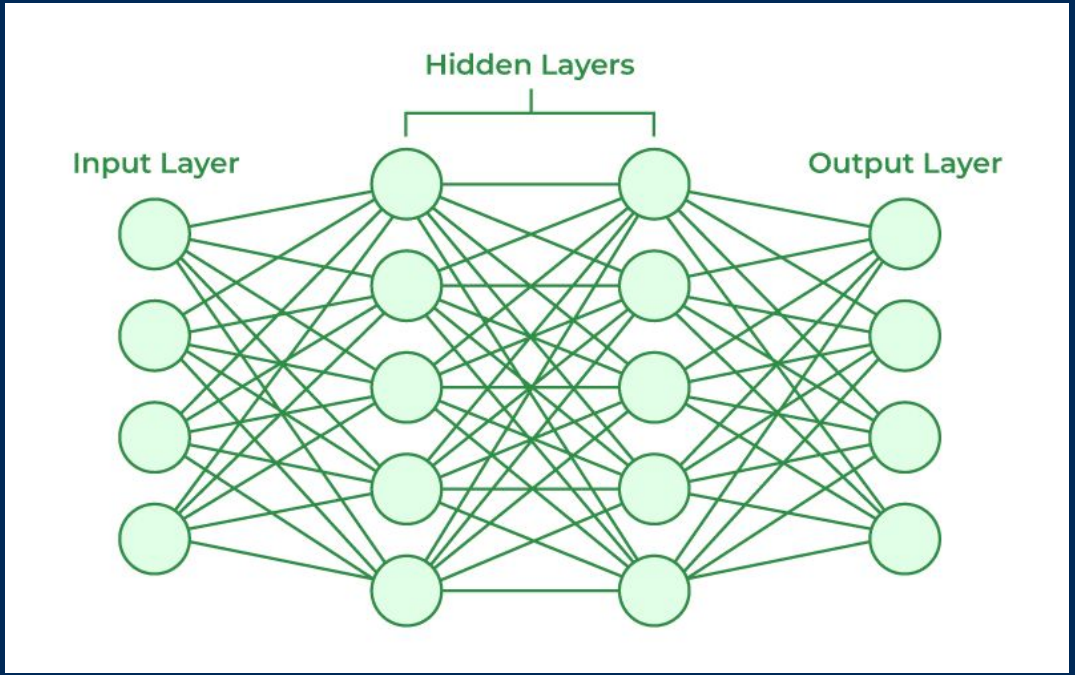



Fig 4. Neural Networks Architecture [8]

# Recurrent Neural Network

Recurrent Unit

Hidden States

- vector h_t

Sequential Data

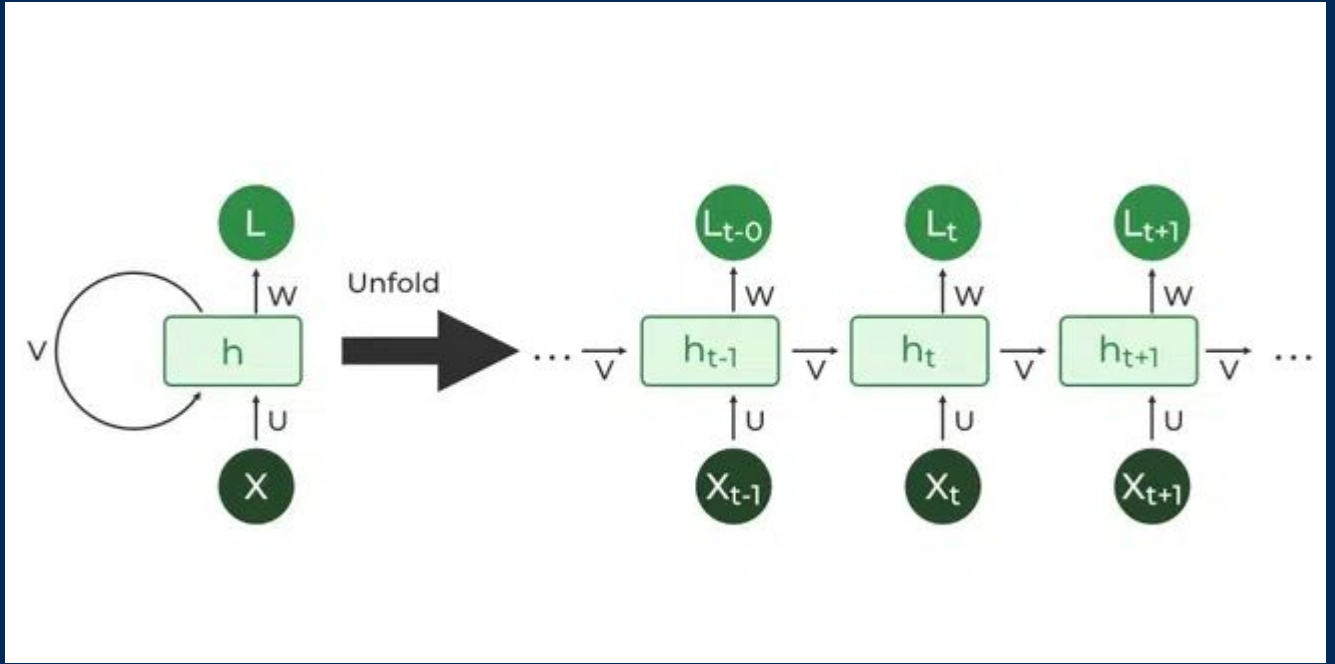



Fig 5. Recurrent Neural Network [9]

# WHAT IS SVM?

- **Definition of SVM**:
  - Supervised machine learning algorithm.
  - Predominantly used for classification tasks.

- **Historical Context**:
  - Developed in 1995 at AT&T Bell Laboratories.
  - Based on the statistical learning framework or VC theory by Vapnik and Chervonenkis.

- **Core Concept**:
  - Seeks a hyperplane that best separates classes.
  - Support vectors are the nearest data points to the hyperplane.
  - Goal is to maximize the margin between these points.

# Mathematical Foundation

- **Objective of SVM**:
  - Find a hyperplane that distinctly classifies the data points.
- **Mathematical Model**:
  - Optimize min **1/2 * ||w||^2**
  - Subject to constraints

  **y_i * (w * x_i + b) >= 1** for each data point
- **Kernel Trick**:
  - Transforms input space to a higher dimensional space.
  - Common kernels: Linear, Polynomial, RBF.

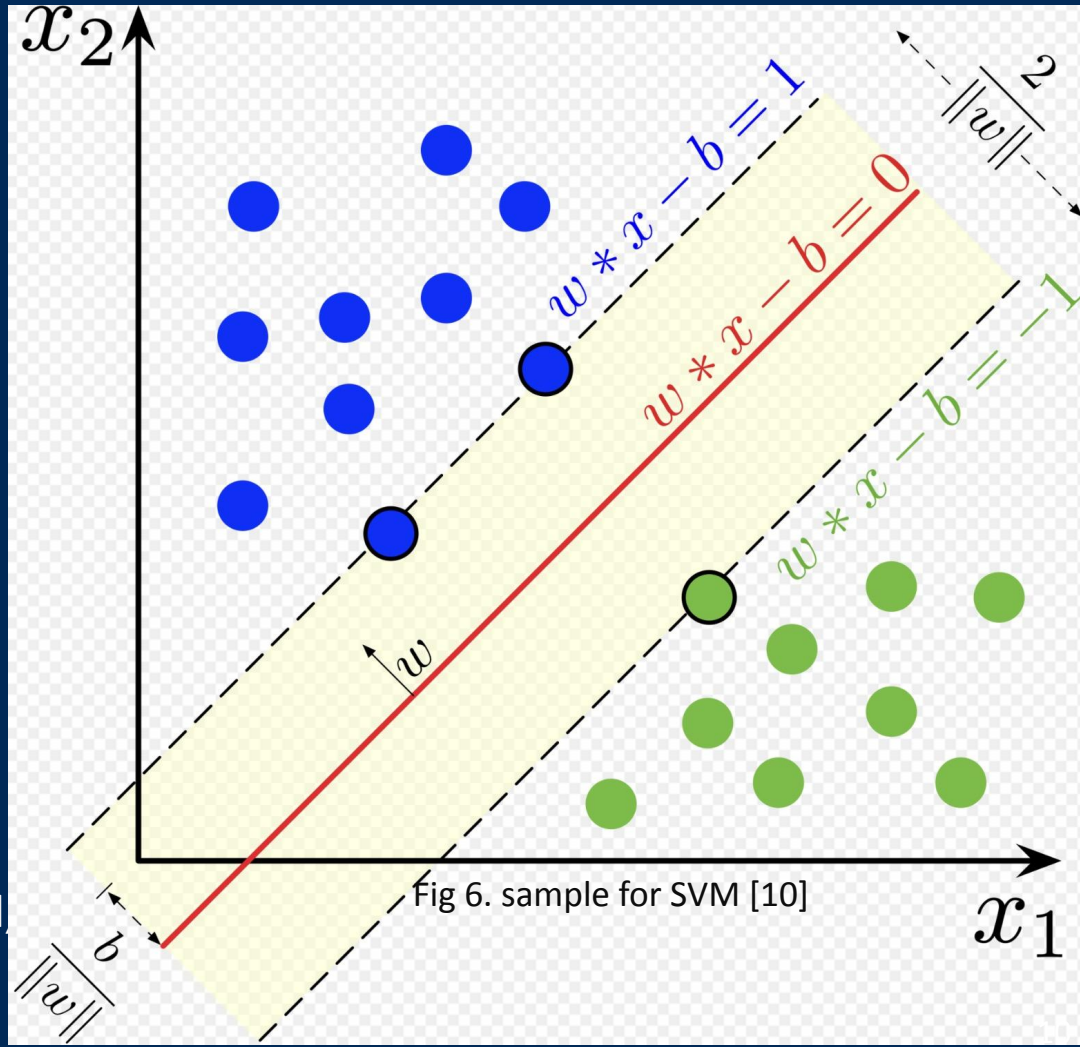



Fig 6. sample for SVM [10]

# Applications of SVM

- **Classification Tasks**:
  - Face detection, image classification, text categorization.
  - Example: Boundary creation around faces in images.
- **Regression Tasks**:
  - Known as SVR.
  - Used for continuous value predictions.
  - Suitable for large-scale regression problems.
- **Industry Applications**:
  - Bioinformatics (protein and cancer classification).
  - Financial sector (credit scoring).

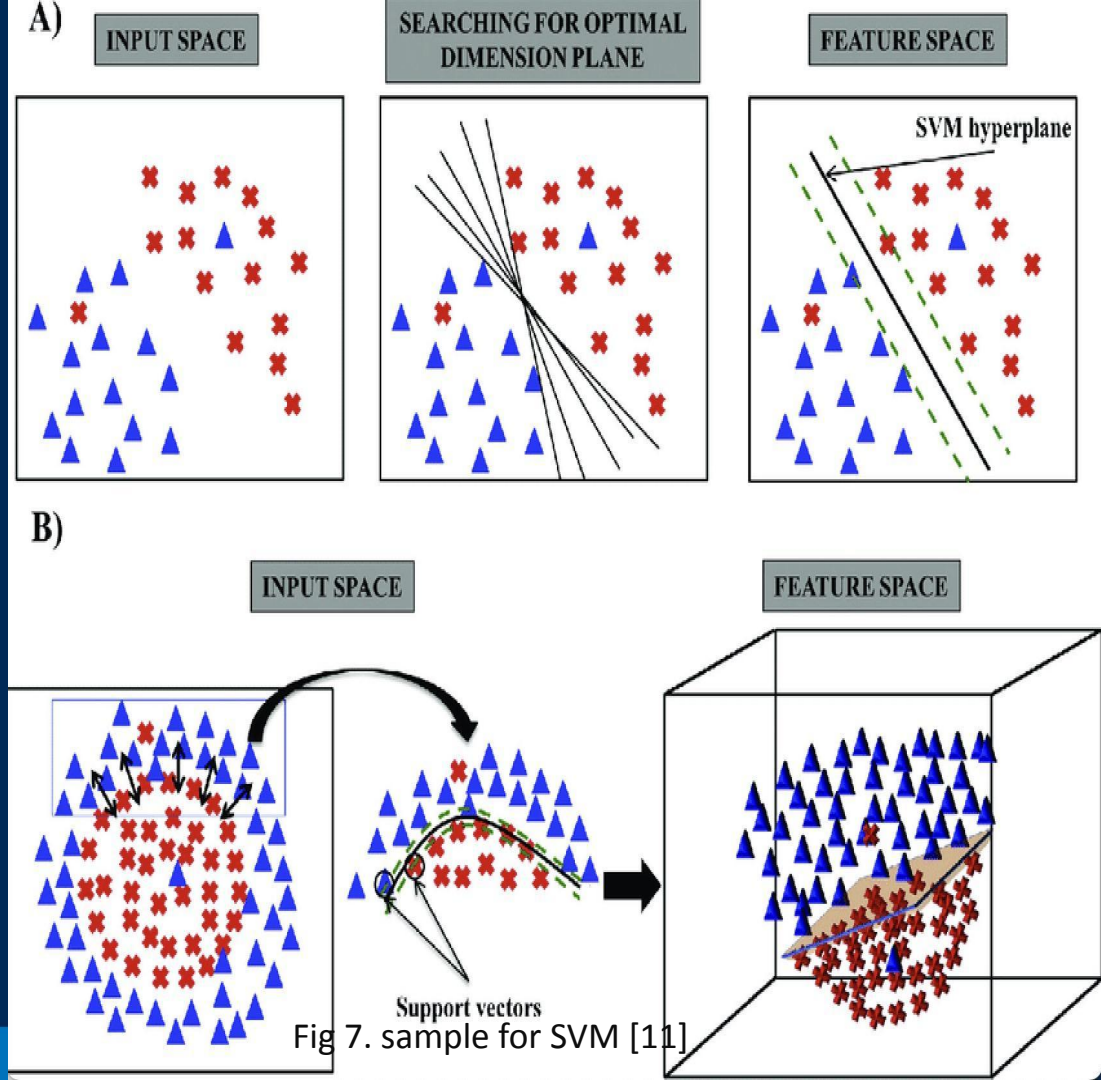



Fig 7. sample for SVM [11]

# Why Choose SVM?

High Accuracy:

Offers excellent accuracy with an appropriate kernel and can be highly effective in high-dimensional spaces.

Effective in High Dimensional Spaces:

Capable of handling very large feature spaces and can perform well even when the number of dimensions exceeds the number of samples.

Versatility in Kernel Choice:

Flexibility to choose from various kernels (linear, polynomial, RBF) or customize your own kernel for the decision function.

# Conclusion

Table 1. Comparison of accuracies of Machine Learning algorithms

| ML Algorithm | Old Result Accuracy | New Result Accuracy (improved using lexical feature analysis on each algorithm) |
|---|---|---|
| Random Forest | 87.34% | 97.369% |
| Support Vector Machine | 89.63% | 97.451% |
| Neural Network with Backpropagation | 89.84% | 97.259% |

*Phishing Detection using Random Forest, SVM and Neural Network with Backpropagation*

With proper feature extraction,

their accuracy could be very close.

# References


[1] What is email phishing and how to prevent it. Brave River Solutions. (2018, January 19). https://braveriver.com/blog/what-is-email-phishing/

[2]What is Random Forest?. IBM. (n.d.). https://www.ibm.com/topics/random-forest#:~:text=Random%20forest%20is%20a%20commonly,both%20classification%20and%20regression%20problems.

[3] Lecture #15: Regression Trees & random forests. (n.d.). https://harvard-iacs.github.io/2017-CS109A/lectures/lecture15/presentation/lecture15_RandomForest.pdf

[4] Classification and detection of email phishing using random ... (n.d.-a). https://norma.ncirl.ie/5126/1/akshatshah.pdf

[5] Akinyelu, A. A., & Adewumi, A. O. (n.d.). Classification of phishing email using Random Forest Machine Learning Technique. Project Euclid. https://projecteuclid.org/journals/journal-of-applied-mathematics/volume-2014/issue-SI16/Classification-of-Phishing-Email-Using-Random-Forest-Machine-Learning-Technique/10.1155/2014/425731.full

[6]Iajit. (n.d.-b). http://iajit.org/PDF/September 2018, No. 5/10600.pdf

[7] AIML.com. (2023a, October 3). What are the advantages and disadvantages of Random Forest? https://aiml.com/what-are-the-advantages-and-disadvantages-of-random-forest/

[8] GfG, “Artificial Neural Networks and its applications,” GeeksforGeeks, https://www.geeksforgeeks.org/artificial-neural-networks-and-its-applications/ (accessed Apr. 17, 2024).

[9] GfG. (2023b, December 4). Introduction to recurrent neural network. GeeksforGeeks. https://www.geeksforgeeks.org/introduction-to-recurrent-neural-network/

[10]*Support Vector Machines*. scikit. (n.d.). https://scikit-learn.org/stable/modules/svm.html

[11] Medium. (n.d.). https://medium.com/low-code-for-advanced-data-science/support-vector-machines-svm-