The explainability of a black-box malware detection model using tabular and NLP-based methods

Junel Alje B. Isanan and Rachel Edita O. Roxas

Abstract— **The exponential surge in malware types, from 100 million in 2012 to 700 million in 2018, poses a significant threat to the cybersecurity sector. While deep learning models have demonstrated impressive accuracy in malware detection and classification, even exceeding 99% in recent studies, the underlying features driving these predictions remain elusive. This lack of transparency hinders user trust in the models' decisions. To address this, this research explores the use of XAI tools to analyze the inner workings of black-box malware detection models trained on tabular and NLP-based data from the EMBER dataset. By uncovering the influential features in single data points, this study aims to enhance model transparency, foster trust in black-box malware detection predictions, and pave the way for improved detection methods and more effective threat analysis in the ever-evolving landscape of cybersecurity.**

Index Terms— **AI, malware, data science, malware data science, malware analysis, XAI, explainable AI, cybersecurity, deep learning, MLP, neural network, black-box models**

I. INTRODUCTION

A. Background of the Study

Malware or "malicious software" is an umbrella term that describes any malicious program or code that is harmful to systems when computers execute [1]. Malware, when attached to computer systems, can bring about serious damage to computer systems necessary for human advancement and conveniences such as registration systems, queueing systems, or even automated sensing systems, among many others. Report shows that 560,000 pieces of malware are created every day, wreaking havoc on personal and company computers alike [2].

Just in recent years, there has been a surge of public interest in the field of artificial intelligence. Tools like ChatGPT, Google Bard, and Midjourney have been constantly utilized by the general public. The mechanism of how AI tools work, particularly AI that uses machine learning, involves learning based on observations made on the training data [3]. Recent studies in the field of cybersecurity machine learning shows that there has been malware detection and classification models that reach near 100% which leaves very little room for research in terms of improving the model accuracy for malware detection models.

However, tabular and NLP-based methods may be crucial for understanding malware detection models in general [4]

especially black-box models such as multi-layer perceptrons (MLP) whose processing mechanisms are difficult to understand and which results are more difficult to interpret in general [5].

B. Statement of the Problem

There is an ever-growing threat of malware, encompassing viruses and Trojans, among many others, posing a range of damaging consequences to individual computers, corporate networks, and the data they contain. The extent of the damage varies from subtle disruptions, where users may remain unaware, to catastrophic failures leading to network breakdowns, data loss, or even theft. These challenges people face in the field of the software industry.

According to a report [6], it is estimated that the financial damages brought by ransomware alone would reach up to $250 billion by 2031.

Another report [7] states that malware is the main cause of the loss of data integrity. This includes data loss and theft. The damage would be magnified if the data was stolen and/or encrypted, especially if it involved personal information or intellectual property.

In recent years, applying machine learning in the field of malware detection has reached 98% to 99% accuracy rates using black-box models [8] [9]. However, explainability of results in black-box malware detection models are yet to be explored fully to gain user trust and improve accuracy of the model [10].

Thereby, this research seeks to address the gap in this field by exploring tabular and NLP-based methods to improve the explainability of black-box models for malware detection, while maintaining or even enhancing their accuracy.

C. Objectives of the Study

This study aims to use an XAI approach in analyzing black-box models.

Specifically, it aims to:
1) Train machine learning models from the data
2) Evaluate model accuracy as secondary methods for establishing user trust in black-box models

3) Use XAI tools to produce feature visualizations of individual predictions
4) Generate counterfactual information on some predictions to assess and verify sensitivity.

D. Significance of the Study

Malware continues to plague us in our day-to-day surfing of the cyberspace.

The analysis of malware data with black-box models allows the malware to be identified immediately upon retraining should a new strain of malware with similar characteristics appear. Thereby, identification of zero-day malware does not need to be constantly searched by cybersecurity experts to calculate heuristics constantly [4].

But as black-box models reach near-perfect accuracy in detecting, with one model reaching 99.8% [9]. There must be a clear shift in priority of researching from improving model to exploring explainability methods to establish user trust in the model's predictions as it is in finance and healthcare [11]. The structural (table) and natural language features of the EMBER dataset are used as a starting point to further show the explainability of a black-box malware model as the data is accessible via thorough cleaning.

II. REVIEW OF RELATED LITERATURE

Today marks an era of modern technology with the widespread accessibility of mobile phones, laptops, and PCs. Technologies have advanced to generate various content, including synthetic personas and prefabricated essays on diverse topics. The convergence of technology and data has allowed mankind to emulate reality and intelligence.

Despite these advancements, questions about the safety of our data and the security of programs persist [12]. The historical encounter with the ILOVEYOU virus underscores the evolution of destructive programs, collectively known as malware.

A. Malware: Purpose, Functions, and Types

Malware, a portmanteau for "malicious software," is a type of software that steals sensitive information and damages or destroys computers/systems [13]. The purposes for creating malware include intelligence and intrusion, disruption, extortion, destruction or vandalism, unauthorized use of computer resources, and monetary gain [13].

B. Data Science

Labelled as the "sexiest job of the 21st century" [14], data science has garnered significant attention. Cielen et al. [15] define data science as the use of methods to analyze large volumes of data and extract meaningful insights. On the other hand, Kudande et al. [16] view it as a scientific discipline that applies scientific methods, algorithms, systems, and processes to derive valuable information from various structured and unstructured data sources.

Cielen et al. [15] outline a typical data science process:

1) **Setting the research goal:** Defining the objectives and questions the analysis aims to answer.
2) **Retrieving:** Gathering the relevant data from various sources.
3) **Data preparation:** Cleaning, transforming, and organizing the data for analysis.
4) **Data exploration:** Analyzing and visualizing the data to uncover patterns, trends, and relationships.
5) **Data modeling:** Developing predictive or explanatory models to gain deeper insights.
6) **Presentation and automation:** Communicating findings and potentially automating the data science process.

This process, as emphasized by Cielen et al. [15], is iterative. It often involves revisiting earlier steps to address challenges, refine data, or improve results. This study will further elaborate on these steps in the methodology section, drawing upon insights from additional research and textbooks.

C. Facets of Data Science

Cielen et al. [15] states that there are multiple facets of data. These are:

- Structured
- Unstructured
- Natural language
- Machine-generated
- Graph-based
- Audio, image, and video
- Streaming

It is important to pick the right facet of data according to the research problem as each facet requires different treatments for preprocessing, modeling, and measuring data. In the case of tabular data, cleaning data would mean imputing based on the mean, median, and mode of column values [17] whereas text data would be cleaned by tokenizing, removing stop words, and stemming [18].

D. The Curse of Dimensionality

Ullah [19] defines high-dimensional data as data with a large number of features, often represented as columns in a tabular dataset. Working with such data presents several challenges:

- **Computational Complexity:** Analyzing high-dimensional data can be computationally expensive. The distance required to search for n nearest neighbors increases with each additional dimension.
- **Limited Human Visualization:** High-dimensional data does not lend itself well to human visualization, making it difficult to analyze the underlying phenomena visually.
- **Redundancy and Noise:** As the number of features increases, so does the potential for redundancy, which introduces noise into the dataset.
- **Exponential Search Space:** The search space within the dataset grows exponentially with added features, causing some algorithms to become unstable.

E. Machine Learning

Machine learning is a fusion of concepts from statistics, mathematics, and numerical analysis to gain insights from data, enabling us to summarize, visualize, group, or predict phenomena within a given dataset [20].

According to Abu-Mostafa et al. [21] in their book *Learning From Data: A Short Course*, there are three types of machine learning:

1) **Supervised Learning:** This technique involves mapping an input to an output, where the output is solely determined by the input. Supervised learning finds extensive use in computer vision, with digit recognition systems trained on the MNIST dataset being a prime example.

2) **Reinforcement Learning:** This technique considers both the player input and the environment state to assess whether transitioning to a new state is advantageous or not. Chess engines [22] exemplify reinforcement learning by grading moves based on user input and the current game state, indicating which player has an advantage and by how much.

3) **Unsupervised Learning:** In this setting, the dataset relies solely on input without any output information. Data points learn independently by discovering patterns and structures within the dataset. An example of unsupervised learning can be seen in market basket analysis, where items are associated based on their frequency of co-occurrence.

F. Multilayer Perceptrons

Multi-layer Perceptrons (MLPs) are a form of feedforward neural network commonly used in supervised learning tasks. They are characterized by three distinct types of densely-connected layers: input, output, and hidden layers serving as the computational core. The output of an MLP is determined depending on the various activation functions utilized (ReLU, Tanh, etc.). Furthermore, the multi-layer perceptron is considered to be a black-box model due to its complex architecture and mysterious decision-making capabilities Dobson$_2$023.

G. Explainable AI

Over the past few years, there is exponential growth in AI production and research with the rise of enormous data and blazing fast computing power. That being said, black-box models can be trained for up to 99+% for a reasonable amount of time given those two conditions earlier. However, black-box models lack explainability given that the values given are only the weights without any transparent decision-making process made by white-box models. Hence, XAI on black-box models rely heavily on approximation and mathematical nuances to calculate probable contribution of features. Below are the some of the listed metrics for XAI [23]:

TABLE I
EVALUATION METRICS FOR EXPLANATIONS

Metric	Description
Sensitivity	The degree to which the explanation is affected by insignificant perturbations from the test point.
Faithfulness	Precisely computes which feature has the most impact on the model output when individually changed.
Simplicity	The ability to choose only the necessary and sufficient features for explaining the prediction.

H. Shapley Additive Explanations

SHAP or SHapley Additive exPlanations is mathematically based on game theory, the Shapley value of a feature is the average expected marginal contribution to the model's decision after all combinations are, however, these values are made possible to be calculated by assuming that the features are independent of each other [24].

III. MATERIALS AND METHODS

This section presents the methodology workflow used for this study as presented in Figure 1. The host machine, development tools, and dataset used for this study are discussed in detail. Initial stages include data cleaning and data preprocessing in preparation for training the black-box malware detection model. After training, the XAI methodology and visualization are outlined.
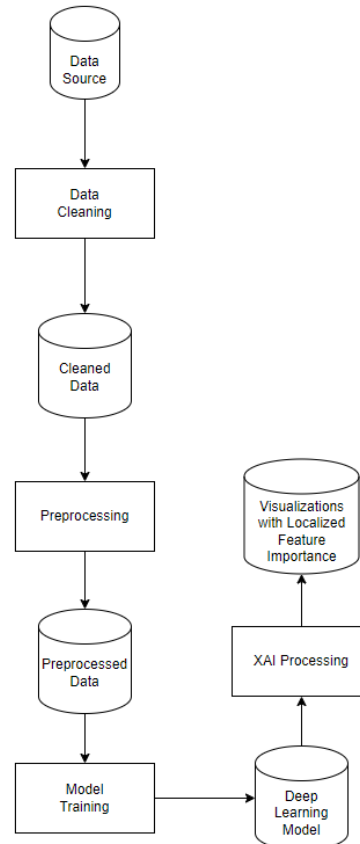

Fig. 1. Methodological Workflow of the Study

A. Host Machine

- **Operating System:** Windows 10 Pro 64-bit (10.0, Build 19045) (19041.vb_release.191206-1406)
- **System Manufacturer:** Colorful Technology And Development Co.,LTD
- **System Model:** H410M-T PRO
- **Processor:** Intel(R) Core(TM) i3-10100F CPU @ 3.60GHz (8 CPUs), 3.6GHz
- **PC Memory:** 16 GB
- **Google Colab Pro Memory:** 52 GB

B. Development Tools

- **Jupyter Notebook:**
  A modular computing environment/text editor where machine learning frameworks / data science related stuff

- **Git:**
  This software provides version control, it is responsible for releasing which versions of the project is stable.

- **Python:** The primary programming language used for developing machine learning models. The language is embedded within Jupyter Notebook.

C. Dataset Used

The dataset used for conducting this study was the EMBER dataset. The EMBER dataset provides eight groups of stringified features that allow for static analysis [25]. These features are categorized as follows:

TABLE II

FEATURES IN MALWARE DATASET

Feature Group	Description
sha256 (id)	Unique identifier for each sample
appeared	Date of appearance
label	Target variable (malicious/benign)
general	*(Discarded)*
header	*(Discarded)*
section	*(Section information about the executable)*
imports	*(Import functions used by the executable)*
exports	*(Discarded)*
histogram	*(Discarded)*
byteentropy	*(Discarded)*
strings	*(String information of the executable)*

For the purposes of this study, the features **section**, **imports**, and **strings** were primarily considered as they were tabular and natural language in nature saxe2018malwaredatascience.

D. Data Cleaning

The feature columns header, imports, and strings came as stringified JSON objects. The study used Jupyter Notebook, pandas, and the JSON library in Python to parse and clean the dirty data in order for these source dataset to become useful.

Retrieval of section names in the were limited as some as the values in the section were obfuscated or packed. The technique was to obtain the usual entry points and section names. Such examples would be .text or .data section. For which code and variable names are written on those sections.

E. Data Preprocessing

Imports were treated as natural language and was therefore tokenized according to frequency using CountVectorizer. The output is a tabularized feature of keywords. Other features are tabularized already, so normalization of values were done for features in Table II.

Due to memory and time limitations, only the top 500 out of the 200k+ most-frequently called imports were utilized.

The basis for cutting down to 500 imports was term frequency of each keywords in order to maximize information capture and at the same time reduce dimensionality for training to be more feasible and faster. [26]. Below are the training accuracies gathered after training some varied number of imports.

TABLE III

PERFORMANCE ACCORDING TO IMPORT SIZE

Import Size	Training Accuracy
500	88.02%
250	86.47%
100	80.58%

Based on the given values, 500 imports would yield better results than the other sizes.

The features are converted using one-hot encoding since they are categorical in nature. This is to ensure that each unique values are represented numerically and are understood by the model.

F. Machine Learning Models Used

The study used a multilayer perceptron (MLP) in order to allow for a thorough learning of pattern in between datasets [27]. An n number of nodes in the input layer is implemented, followed by 64 nodes of hidden layer and lastly, 2 nodes for the output layer, classifying benign and malicious probability values. The corresponding activation functions used for the hidden layer were ReLU and sigmoid was implemented for the output layer. It is the purpose of the study to use black-box models such as multi-layer perceptrons to extract insights due to their high accuracy [21] but opaque decision-making methodology [5]

G. XAI Methodology

The XAI Methodology used in this study is Shapley Additive Explanations (SHAP) and the prediction is displayed on a bar chart in order of descending individual feature importance. SHAP cannot attribute the totality of the feature on a global scale. The prediction is done post-hoc, meaning that the

TABLE IV

MLP ARCHITECTURE FOR MALWARE DETECTION

Layer Type	Configuration	Output Shape
DenseFeatures	Depending on the size of feature	(Num. Features)
Dense	64 neurons, ReLU activation	(64)
Dropout	0.2 dropout rate	(64)
Dense	2 neurons, Sigmoid activation	(2)

prediction will be done after the model has been trained and is fed a number of individual data points for features to be predicted.

IV. RESULTS AND DISCUSSION

This section addresses the objectives of the study and is subdivided into three parts:

1) Model Training and Accuracy Evaluation
2) Visualization of Features
3) Counterfactual Generation
4) Other Predictions

A. Model Training and Accuracy Evaluation

On Table V, the post-hoc explainability results of the model is shown. This means that the explanation comes only after a single datapoint has been predicted.

TABLE V

MODEL PERFORMANCE AND EXPLAINABILITY BY FEATURE

Feature	Model Used	Accuracy	XAI Method
Imports	MLP	88.02%	SHAP
String	MLP	84.26%	SHAP
Header	MLP	76.04%	SHAP

Since the study focuses on explainability, Table V is merely for analytics purposes. Most of the study results will delve into the explainability of individual predictions.

B. Visualization of Features

This subsection presents the visualizations of the XAI model. Visualizations are presented in a descending order based on magnitude of local feature importance.

To interpret the values on each visualization, features leaning to the orange color imply that these features contribute to the prediction of a single data point malware. On the other hand, features leaning to the blue color imply that these features contribute to the benignity of the data point. The longer the bar, the greater the share of the feature's prediction towards a certain value.

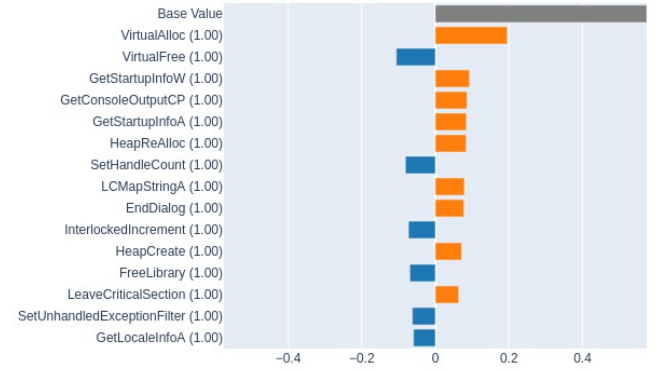A sample data point predicted using the imports feature as shown in Figure 2.

Fig. 2. Sample data point predicted using the import feature

In Figure 2, it is shown that 18.5% of the prediction has been attributed to VirtualAlloc alone. It implies that on the absence of this single feature, the prediction of this malware will change drastically.

C. Counterfactual Generation

In this subsection, a counterfactual generation data point is created to test the sensitivity and faithfulness of the data. When the most salient feature is removed, it should significantly lower the prediction of the data point by the model.

Based on Figure 2, it is 92.6% certain that the data point is malware. On Figure 3, however, the absence of VirtualAlloc is simulated while the other features remain intact. As a result, the prediction drastically changes to 40.6% as evidenced in Figure 3.
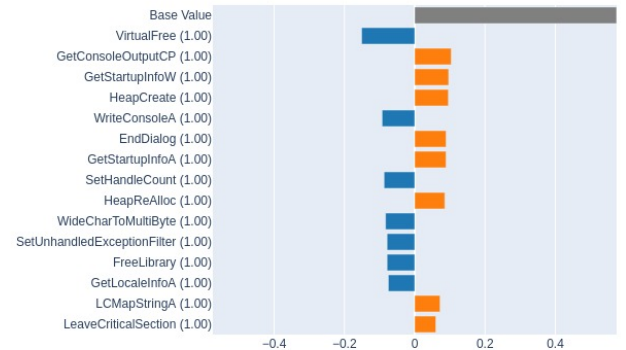

Fig. 3. The same data point in Fig. 2 after counterfactual information was generated. VirtualAlloc was removed from the datapoint

By changing just one of the most important features of the local prediction, not only did it have a drastic change in prediction, but also the model predicts benignity in the code despite the ground truth stating otherwise.

D. Other Predictions

Regarding other predictions, Figures 4 and 5 are samples of individual predictions for other features.

As shown on Figure 4, the prediction implies that most of the features seem to connive that the said data point is malicious based on the given graph.

On the other hand, Figure 5 doesn't seem to be coherent at all since it violates the XAI principle of simplicity. Each feature should have semantically coherent implications.

Furthermore, the study also produced 10 predictions per feature as a way to explain other local predictions.
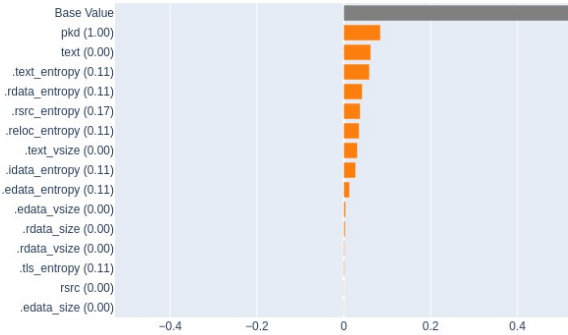
Fig. 4. Another datapoint predicted using the headers feature and discovered it was a packed malware (due to the pkd feature which was custom-built for the study)
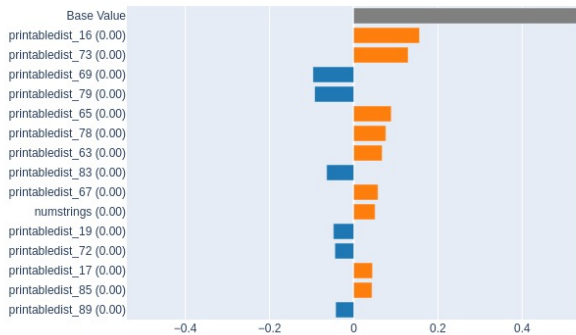

Fig. 5. Another datapoint predicted using the strings feature, but it doesn't seem to be coherent at all

V. CONCLUSION AND FUTURE WORK

With this study, black-box malware detection models were interpreted locally using SHAP. This will allow cybersecurity experts to comment and improve on existing machine learning models.

As for the future works, with Kolmogorov-Arnold Networks (KAN) as the emerging trend for deep learning, where activation functions are placed on edges instead of nodes. This type of neural network may offer more explainability due to the individual weights placed on edges instead of nodes [28]. For the meantime, while the capabilities of Kolmogorov-Arnold Networks are not yet fully researched, the study could be enriched further with neural networks with the help of domain experts to determine whether such features predicted by the model using game-theoretic approaches are relevant using counterfactual generation.

Moreover, the explainability of machine learning models can easily be transferred to other sectors such as healthcare, agriculture, social sciences, and humanities. XAI will help machine learnings improve by counterchecking on features/pixels that influenced such decisions. The source code can easily be tweaked according to need.

And lastly, validate the truthfulness of the features assigned by the model with a domain expert.

APPENDIX I
OBFUSCATED SECTION FILES

Fig. 6. Photo of obfuscated sections

Figure 6 represents some of the sections that were obfuscated out of the thousands retrieved from the dataset. Indicating that the section texts might be hidden to prevent from being reverse-engineered.

APPENDIX II
GENERATION OF COUNTERFACTUAL INFORMATION

```
 ✓  [216]  1 sixth_row = X_predict.head(6).tail(1)
Os

 ✓  [221]  1 sixth_row['VirtualAlloc']
Os
    ⇥  5    1
         Name: VirtualAlloc, dtype: int64

    [135]  1 sixth_row['VirtualAlloc'] = 0
```

Fig. 7. Generation of counterfactual information

Figure 7 represents the method on how to remove the most salient feature in Figure 2 which is VirtualAlloc.

REFERENCES

[1] Apr 2024. [Online]. Available: https://www.malwarebytes.com/malware
[2] N. J. Palatty, "30+ malware statistics you need to know in 2024," Dec 2023. [Online]. Available: https://www.getastra.com/blog/security-audit/malware-statistics/
[3] SAS, "Machine learning: What it is and why it matters," n.d. [Online]. Available: https://www.sas.com/en_us/insights/analytics/machine-learning.html
[4] J. Saxe and H. Sanders, *Malware data science: Attack detection and attribution*. No Starch Press, 2018.
[5] J. E. Dobson, "On reading and interpreting black box deep neural networks," *International Journal of Digital Humanities*, vol. 5, no. 2–3, p. 431–449, Nov 2023.
[6] C. Sausaulito, "Global ransomware damage costs predicted to exceed $265 billion by 2031," *Cybercrime Magazine*, 2023. [Online]. Available: https://tinyurl.com/bdcppzhy
[7] K. I. Encyclopedia, "Damage caused by malware," n.d. [Online]. Available: https://encyclopedia.kaspersky.com/knowledge/damage-caused-by-malware/
[8] G. M. and S. C. Sethuraman, "A comprehensive survey on deep learning based malware detection techniques," *Computer Science Review*, vol. 47, p. 100529, Feb 2023.
[9] A. K. Yuksel and Y. Ar, "A machine learning approach to malware detection using application programming interface calls (mdapi)," *Traitement du Signal*, vol. 40, no. 4, p. 1511–1520, Aug 2023.
[10] M. Li, H. Sun, Y. Huang, and H. Chen, "Shapley value: From cooperative game to explainable artificial intelligence," *Autonomous Intelligent Systems*, vol. 4, no. 1, Feb 2024.
[11] Y. Liu, C. Tantithamthavorn, L. Li, and Y. Liu, "Explainable ai for android malware detection: Towards understanding why the models perform so well?" *2022 IEEE 33rd International Symposium on Software Reliability Engineering (ISSRE)*, Oct 2022.
[12] C. P. Software, "Biggest cyber security challenges in 2023," *Check Point Software*, 2023. [Online]. Available: https://www.checkpoint.com/cyber-hub/cyber-security/what-is-cybersecurity/biggest-cyber-security-challenges-in-2023/
[13] Cisco. (2023, Jul) What is malware? - definition and examples. [Online]. Available: https://www.cisco.com/site/us/en/products/security/what-is-malware.html
[14] T. H. Davenport and D. Patil, "Data scientist: The sexiest job of the 21st century," Mar 2024. [Online]. Available: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century
[15] D. Cielen, M. A. D. B., and A. Mohamed, *Introducing data science: Big data, Machine Learning, and more, using python tools*. Manning, 2016.
[16] P. D. Kudande, "A review paper on the formation of data science," *International Journal of Engineering Applied Sciences and Technology*, 2020. [Online]. Available: https://tinyurl.com/muxs43ud
[17] C. Fan, M. Chen, X. Wang, J. Wang, and B. Huang, "A review on data preprocessing techniques toward efficient and reliable knowledge discovery from building operational data," *Frontiers in Energy Research*, vol. 9, Mar 2021.
[18] V. Gurusamy and S. Kannan, "Preprocessing techniques for text mining," 10 2014.
[19] F. Ullah, 2019. [Online]. Available: https://web.lums.edu.pk/\imdad/pdfs/CS5312_Notes/CS5312_Notes-11-Curse_of_Dimensionality.pdf
[20] W. J. Deuschle, "Undergraduate fundamentals of machine learning," 2019.
[21] Y. S. Abu-Mostafa, M. Magdon-Ismail, and H.-T. Lin, *Learning from data: A short course*. AMLBook.com, 2012.
[22] S. Maharaj, N. Polson, and A. Turk, "Chess ai: Competing paradigms for machine intelligence," *Entropy*, vol. 24, no. 4, p. 550, Apr 2022.
[23] L. Coroama and A. Groza, "Evaluation metrics in explainable artificial intelligence (xai)," *Communications in Computer and Information Science*, p. 401–413, 2022.
[24] V. Belle and I. Papantonis, "Principles and practice of explainable machine learning," *Frontiers in Big Data*, vol. 4, Jul 2021.
[25] H. S. Anderson and P. Roth, "Ember: An open dataset for training static pe malware machine learning models," 2018.
[26] K. Ganesan, "What is term frequency?" Aug 2021. [Online]. Available: https://www.opinosis-analytics.com/knowledge-base/term-frequency-explained/#:~:text=Term%20frequency%20(TF)%20means%20how,about%20how%20you%20define%20it.
[27] L. Alzubaidi, J. Zhang, A. J. Humaidi, A. Al-Dujaili, Y. Duan, O. Al-Shamma, J. Santamaría, M. A. Fadhel, M. Al-Amidie, and L. Farhan, "Review of deep learning: Concepts, cnn architectures, challenges, applications, future directions," *Journal of Big Data*, vol. 8, no. 1, Mar 2021.
[28] C. J. Vaca-Rubio, L. Blanco, R. Pereira, and M. Caus, "Kolmogorov-arnold networks (kans) for time series analysis," 2024.

Junel Alje B. Isanan is a BS Computer Science student at the University of the Philippines Los Baños. Hailing from Polomolok, South Cotabato, his passion for technology and its potential for positive impact led him to pursue data science. As the current Logistics Co-Director of the UP Data Science Society (AY 2023-2024), Junel actively fosters a community of data enthusiasts. His leadership experience also includes serving as the Scholastics Department Head of the Young Software Engineers' Society (AY 2022-2023), where he taught certain lessons among aspiring software engineers. Beyond his technical pursuits, Junel finds joy in exploring new places and broadening his horizons.

Rachel Edita O. Roxas is a Professor at the Institute of Computer Science at the University of the Philippines Los Baños. She has been awarded as National member finalist representing the Philippines in the Senior Scientist category of the ASEAN Women in Science 2022 Awards (organized by ASEAN through COSTI, USAID, and the Underwriters' Laboratories), and named among the top 1,000 scientists in the Philippines by the AD Scientific Index consistently over the past years, while focusing on Natural Language Processing and its applications towards participative governance.