Jurnal Ilmiah Teknik Elektro Komputer dan Informatika (JITEKI)
Vol. 10, No. 4, December 2024, pp. 747-762
ISSN: 2338-3070, DOI: 10.26555/jiteki.v10i4.30157

Using Graph Neural Networks and CatBoost for Internet Security Prediction with SMOTE

Aswan Supriyadi Sunge[1], Spits Warnars Harco Leslie Hendric[2], Dendy K. Pramudito[1]
[1]Informatics Engineering Department, Pelita Bangsa University, Bekasi, Indonesia
[2]Computer Science Department, Graduate Program-Doctor of Computer Science, Bina Nusantara University, Jakarta, Indonesia

ARTICLE INFO

Article history:

Received September 26, 2024
Revised December 01, 2024
Published December 20, 2024

Keywords:

Predictions;
Website;
Security;
CatBoost;
GNNs

ABSTRACT

Internet security is the most important issue in cyberspace, on the other hand, cybercrime occurs, and the most serious threat is the theft of personal data and its misuse for the benefit of others. Although cyberspace is while internet security cannot eliminate all risks, predictive models can significantly reduce cybercrime by identifying vulnerabilities if you know how to prevent it. One of the most important things is that many internet users do not know what measures are used to avoid and whether it is safe to visit or explore, on the other hand, in system development existing studies on internet security prediction often rely on generic models that lack precision in identifying influential features or ensuring class balance in developing internet security. In this case, Deep Learning (DL) helps learn patterns from recorded data, find relevant patterns, and use the model effectively. The purpose of this study is to identify the most influential features in internet security and evaluate the effectiveness of advanced machine learning models, such as Graph Neural Networks (GNNs) and Categorical Boosting (CatBoost), for predicting internet safety. So far other studies have tested the entire data set and used a model that is generally. This is expected to lead to the design or development of systems and programs that are useful for internet security. The study used a dataset of 11,055 records with 30 features and binary classification labels ('Safe' and 'Not Safe'). To address the class imbalance, SMOTE was applied before splitting the data into training and testing sets. In testing the Graph Neural Networks (GNNs) model achieved 93.58% accuracy, 93.63% precision, 93.58% recall, and 93.55% F1-score, demonstrating its effectiveness for internet security prediction. From the results of testing the CatBoost model was used to identify key features, revealing that the 'URL of Anchor,' 'SSLFinal State,' and 'Web Traffic' have the most significant impact. From the experiments conducted, the CatBoost effectively identified features with the highest on prediction accuracy, and the GNNs model is very accurate and precise for developing applications or systems to predict internet security.

Corresponding Author:

Aswan Supriyadi Sunge, Pelita Bangsa University, Inspeksi Kalimalang Street, Bekasi 17530, Indonesia.
Email: aswan.sunge@pelitabangsa.ac.id

1. INTRODUCTION

The Internet is a change in modern communication with a global network of social media services, entertainment, health, education, and so on [1]. The rapid growth of information technology, coupled with increasing internet penetration, has connected a significant portion of the global population, especially since it has entered and almost the entire world's population is connected to the internet in everyday life and is the main fundamental in world infrastructure [2], [3]. However, on the other hand behind the extraordinary opportunities for internet users from personal to privacy and negligence against internet crime [4], [5], [6], especially users are not aware and do not know how to deal with these attacks [7], [8]. However, it is recognized that there are already many methods of preventing internet crime providing awareness with the slogan of healthy internet [9], routinely changing passwords [10], updating operating systems on computers or mobile [11], including applications or systems that prevent internet crime attacks based on IoT and blockchain [12], [13], [14]. Despite advancements in internet security, predicting vulnerabilities remains a challenge due to imbalanced datasets, redundant features, and the lack of optimized models. This study aims to address these gaps using advanced DL models and feature selection techniques

From the technology or research that has been done in preventing crime, it all boils down to the use of pattern recognition or prediction based on datasets using models such as Convolutional Neural Networks (CNNs), Recurrent Neural Networks (RNNs), Graph Attention Network (GAT), Bidirectional Long and Short-term Memory Network (Bi-LSTM), Deep Neural Networks (DNNs), and Long Short-term Memory Network (LSTM) [15]-[19]. These technologies are necessary and used because it is very popular in terms of more efficient performance, complexity, and accuracy [20], and are also very effective in learning representative features from datasets [21]. These models are applied in various fields [22], In other fields such as forest fire mapping [23], potato sorting mapping [24], rice variety identification [25], and regional mapping [26], [27].

However, in addition to the use of various models, exploration is also needed in selecting which features are relevant or influential. This step is needed to improve the learning process effectively, efficiently, and accurately, especially removing irrelevant or redundant features and improving the quality of data analysis [28]-[31]. There are several techniques or models in searching for features that are influential or relevant in data analysis or model development. These include methods such as Pearson Correlation, a method of measuring the level of correlation strength or relationship of each feature with the target feature or class [32], [33], [34]. SelectKBest is a method of selecting the highest features based on statistical tests and is available in the scikit-learn library in the Python programming language [35], [36]. Chi-Square is a method of testing between two variables and analyzing the relationship between them [37], [38], [39]. SHAP stands for SHapley Additive exPlanations, which is an interpretation in finding the highest feature selection [40], [41], [42]. LIME, an abbreviation of Local Interpretable Model-Agnostic Explanations, is a prediction model that displays influential features or attributes and is easy to understand in terms of interpretation and visuals [43], [44], [45].

Feature search models are widely used, such as in the health sector for heart attack prediction [46], [47], [48], care for disabled children [49], cancer [50], [51], [52], and tumors [53]. In the economic sector such as credit risk [54], carbon futures pricing [55], exchange rate prediction [56], in other fields such as energy consumption prediction [57], [58], prediction of electricity usage in the agricultural sector [59], potato disease prediction [60], wind power [61], email spam [62], daily electricity prices [63], network security in LTE/LTE-A [64], wireless sensor network security [65], botnet attacks on IoT [66].

The contribution of this research is the development of an internet security prediction model that combines advanced techniques with the most relevant and influential feature search algorithm. The model used is GNNs, designed to identify complex patterns and relationships in data with a graph structure, thereby increasing the accuracy in predicting potential security threats. In addition, they are optimizing the search for influential features using the CatBoost algorithm, which is known to have excellent capabilities in handling large and imbalanced data and providing more accurate and efficient prediction results. This approach provides a more effective solution in detecting and preventing potential security threats on the internet.

2. SIMILAR PREVIOUS RESEARCH

One of the technological discoveries of this century is the discovery of the internet is one of the greatest innovations and influences all aspects of human life changes from the social, economic, and industrial fields [67]. Along with increasing population growth and supported by the level of education, technology changes are also increasing [68]. In the past when people wanted to meet, they had to meet face to face, now they can do it by phone or video call, shopping had to go to the store, now just order by phone and the goods will be delivered, learning is easier without having to meet face to face, learning can be done anywhere, time can be adjusted and many things are easy when connected to the internet. With the internet, it becomes easier and simpler, but behind the convenience, there are many crimes in cyberspace from data theft, pornography, lack of socializing, and receiving unclear information [69], which ultimately internet users become victims of internet crime and especially personal data is misused which is detrimental to others.

It is indeed recognized that there are already applications or systems that detect internet crimes based on suspicious programs, but that is not enough, so another method is needed with DL. The discovery of advanced techniques provides a new solution to solving problems from computer vision, Natural Language Processing (NLP), and Artificial Intelligence (AI) and shows very good performance. The use of DL is used in several fields such as pedestrian fatigue detection [70], indoor object identification [71], traffic sign identification [72], making applications in medicine [73], and identifying internet crimes in social media [74], phishing which is a technique of stealing personal information without the user realizing it [75], IoT security in Smart City [76].

There are several models one of GNNs, which is one of the Neural Networks techniques and one of the most powerful weapons to display graphs in conveying messages implied in the dataset [77] and can be applied in data mining [78], [79], especially in improving the accuracy of prediction results [80]. Different from other models such as RNNs which focus more on sequential data such as time series, text, speech, and so on [81], or CNNs, which focus more on things like videos and images [82].

The use of GNNs is widely used in city planning in wind prediction [83], passenger demand prediction in vehicle reservations [84], diabetes complication prediction [85], network problem prediction [86], weather prediction [87], drug prediction with the emergence of new drugs [88]. In network security prediction or cybercrime, in malware prediction in the Android system [89], fake video prediction in hoax news [90], Malware prediction, such as ransomware, trojans, spyware, and botnets in the Windows operating system [91] and security prediction in IoT [92], therefore the GNNs model is very suitable in predictions, especially in this study.

However, the use of DL is not enough because it only looks at the prediction accuracy, it is necessary to see which features are influential or highest. One of the best models in feature search is CatBoost which is the latest generation of Machine Learning (ML) which is faster than the previous XGBoost [93] and mainly supports categorical data which is usually only related to numerical data [94], [95], [96], the best compared to other classification models [97] and shows the highest accuracy results and minimal errors in displaying features [98], [99]. The use of CatBoost model is widely used such as in the health sector in cardiovascular prediction which is a disease caused by disorders of the heart and blood vessels [100], and prediction of heart rate conditions because the average human heart rate is 60 bpm but if it exceeds that rate then there is something [101], [102], diabetes which is a metabolic disorder that increases blood glucose levels [103], [104], [105], Cervical cancer is the disease that most commonly attacks adult women in the world [106], stroke is an acute medical disorder where a blood artery in the brain ruptures which results in loss of consciousness [107]. In the economic sector such as predicting financial risks to stocks that are not listed and not traded on the securities market [108], predicting urban development [109], and predicting indexed stock trends [110]. In the field of network security or cybercrime such as predicting scammer identification [111], and predicting phishing websites [112].

By combining both models, it is an effective method to improve prediction accuracy. GNN excels in capturing complex relationship patterns in graph-based data, and CatBoost has the advantage of identifying the most relevant and significant features through a deep approach to categorical data.

3. PROPOSED MODEL

This research is an experimental study using quantitative methods based on the literature review and theoretical framework obtained and then comparing the performance of models in the context of internet security prediction using the GNNs model to measure accuracy, precision, recall, and F1-Score and the CatBoost model to test in finding the highest or influential features. The main objective is to provide new insights and significant contributions to the field of internet security and research questions include performance comparison and feature selection.

Fig. 1 depicts two different analysis processes with two colored lines with specific meanings. The green line represents Research Question 1 (RQ1), which illustrates the testing phase with the GNNs model to evaluate the results for Accuracy, Precision, Recall, and F1-Score. Meanwhile, the red line represents Research Question 2 (RQ2), which depicts the process of searching for or identifying the highest or most influential features using the CatBoost model. Therefore, the green line focuses on the evaluation of the performance metrics of the GNNs model. In contrast, the red line focuses on selecting relevant features via the CatBoost model.

3.1. Start

This research begins with collecting data related to internet security, then looking for relevant and complete information according to the existing problem. The data collected covers various aspects of the internet such as IP addresses, URLs, ports, and so on. Also, ensure this data is accurate and represents different groups of internet users. Next, explain how this data was collected, what features and amount of data it includes, and prepare it to be explained. The goal is to use this information to increase user safety while surfing.

3.2. Datasets

This section is to identify the data used for this study and ensure that the data used can be tested perfectly and is very relevant. The data used in Table 1, downloaded and stored on the UCI (ML) Repository site with the link https://archive.ics.uci.edu/dataset/327/phishing+websites, which is a collection of datasets by the machine learning community to be analyzed with the algorithm to be used, also provided periodically and publicly available, the dataset is from 11055 data, 30 attributes, and 1 class (-1 = Not Safe, 1 = Safe). The purpose of this dataset is to predict whether the website is safe or not when visited.

The dataset reflects the threat to internet security due to various techniques and tactics used by attackers to deceive users and steal personal information. Moreover, techniques or attacks are increasingly developing and more sophisticated, as well as the use of social media techniques to deceive users. However, it is recognized that the limitations of this dataset are the lack of the latest attack tactics, as well as reliance on features that may not always be detected by security systems. In addition, this dataset may not fully represent global variations, because most of the data comes from limited sources or specific regions. Nevertheless, the analysis of this dataset still provides important insights in detecting and preventing internet security.

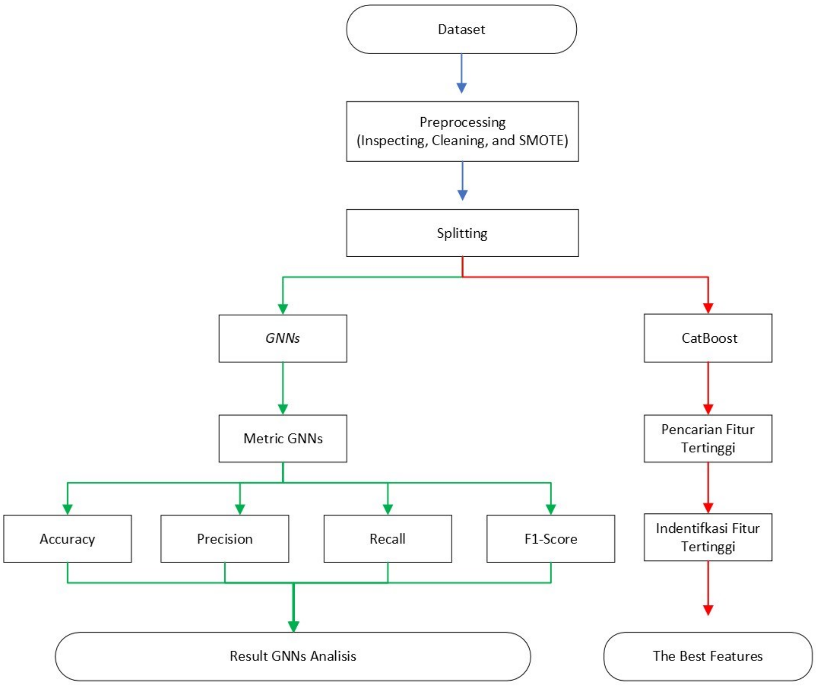

Fig. 1. Framework Model

Data processing begins with the preprocessing stage, this phase prepares the data well for testing to reduce data errors and support the ML data analysis process. There are several phases, depending on the desired results and the model used, but this phase is very important to know the quality of the data before testing. The most common techniques are data checking and cleaning techniques, unclear symbols, duplicates, unclear, typos, or data deletion. The goal is to ensure that the data is filled or not empty, and in particular the data is complete and free from errors. Testing uses the data. *data.isnull(().sum()* command syntax to display the number of errors or missing data and test results in Table 2.

The second phase tests the balance between the number of classes or individual labels in the data being tested. This is intended to determine whether there is an imbalance in the data and whether it produces significantly higher or lower classes or labels. The goal is to see most of the prediction data and can reduce overall performance. For example, in the prediction of this study, there may be more uncertain results than specific data, and vice versa. If this problem occurs, it is necessary to find a solution by cleaning and relocating the data. The test uses the syntax *target1=data[data['Class']==1] and target1=data[data['Class']==0]*. However, the dataset used in this study contains Imbalanced Data, which means that the condition of a class has an unbalanced number or the number of class data is much different compared to other classes. So the Synthetic Minority Over-sampling Technique known as SMOTE is used, a method for handling unbalanced data so that the minor class is balanced with the major class [113].

However, using SMOTE to handle a class imbalance in a dataset can risk overfitting due to synthetic data generation, which can introduce irrelevant or redundant patterns into the model. To overcome this problem, combine it with an ML model that can handle complexity and generalization well, such as Regularization or Cross-Validation. This approach helps reduce the risk of overfitting by ensuring that the model remains relevant and can identify patterns from both the minority and majority classes. The results of the SMOTE test before and after are shown in Fig. 2.

Table 1. Data Set Features and Description

No	Data Set		
	Features	**Description**	**Value**
1	Having_IP_Address	If the device or hardware has an IP address.	-1 or 1
2	URL_Length	Access the resources contained on the web page.	-1 or 1
3	Shortining_Service	A service that shortens URLs.	-1 or 1
4	Having_At_Symbol	There is an "@" symbol in emails or communication formats.	-1 or 1
5	Double_slash_redirecting	A situation where the URL redirects causing two "//" signs to be generated.	-1 or 1
6	Prefix_Suffix	A term used to refer to a string or word.	-1 or 1
7	Having_Sub_Domain	The existence of subdomains within a domain name.	-1 or 1
8	SSLfinal_State	The final stage in the negotiation process in setting up a connection.	-1 or 1
9	Domain_registeration_length	Length of time in domain name registration.	-1 or 1
10	Favicon	Short for "Favorite Icon" which displays a small icon to represent a website.	-1 or 1
11	Port	Refers to identifying and managing communications to various systems.	-1 or 1
12	HTTPS_token	A concept used in the context of secure communications on the internet.	-1 or 1
13	Request_URL	It is a request sent by a client to a server in web communication.	-1 or 1
14	URL_of_Anchor	Elements used to create links that direct users to other websites.	-1 or 1
15	Links_in_tags	Links that contain HTML elements that use tags.	-1 or 1
16	SFH	Short for "Self-Referencing File Host" in the context of web security and programming.	-1 or 1
17	Submitting_to_email	The process of sending data in the form of a formula via email.	-1 or 1
18	Abnormal_URL	It is a characteristic or pattern that does not conform to the standard URL format.	-1 or 1
19	Redirect	A technological method that refers to the process of requesting from a URL to another URL.	-1 or 1
20	On_mouseover	Using the mouse cursor to move upwards or certain elements on the web.	-1 or 1
21	RightClick	A command that clicks more to the right than to the left.	-1 or 1
22	PopUpWindow	It is an automatic window that appears above the main browser window.	-1 or 1
23	Iframe	Web elements used to display other information within the web.	-1 or 1
24	Age_of_domain	Age or length of time the domain has been registered since it was first registered.	-1 or 1
25	DNSRecord	Maps domain names to various types of information related to that domain.	-1 or 1
26	Web_traffic	The number of visitors or users to a website.	-1 or 1
27	Page_Rank	It is an algorithm that determines web ranking in search results.	-1 or 1
28	Google_Index	Websites listed in the Google index.	-1 or 1
29	Links_pointing_to_page	It is the number or links that lead to a particular web page.	-1 or 1
30	Statistical_report	This is a report that presents visitor statistics, how many hours they visited, or other data related to the website.	-1 or 1

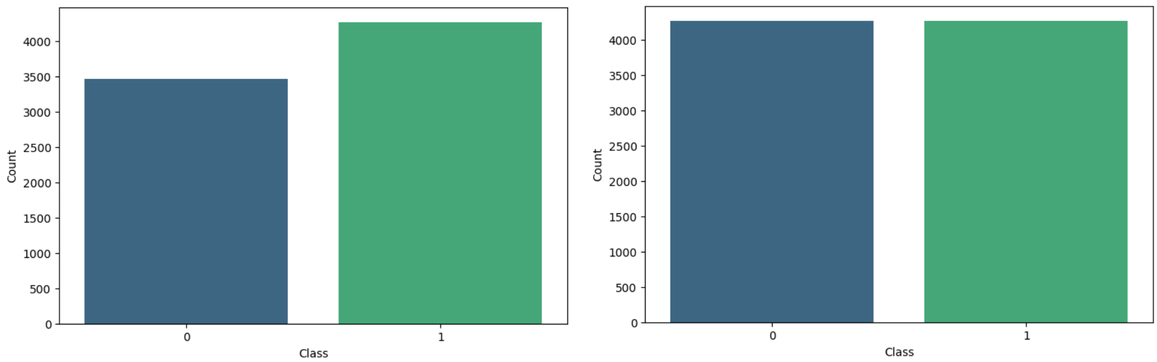

Fig. 2. (a) Class Distribution before SMOTE; (b) Class Distribution after SMOTE

Table 2. Test Results Inspecting and Cleaning data

No	*data.isnull().sum()*	
	Features	**Value**
1	Having_IP_Address	0
2	URL_Length	0
3	Shortining_Service	0
4	Having_At_Symbol	0
5	Double_slash_redirecting	0
6	Prefix_Suffix	0
7	Having_Sub_Domain	0
8	SSLfinal_State	0
9	Domain_registeration_length	0
10	Favicon	0
11	Port	0
12	HTTPS_token	0
13	Request_URL	0
14	URL_of_Anchor	0
15	Links_in_tags	0
16	SFH	0
17	Submitting_to_email	0
18	Abnormal_URL	0
19	Redirect	0
20	On_mouseover	0
21	RightClick	0
22	PopUpWindow	0
23	Iframe	0
24	Age_of_domain	0
25	DNSRecord	0
26	Web_traffic	0
27	Page_Rank	0
28	Google_Index	0
29	Links_pointing_to_page	0
30	Statistical_report	0

3.3. Splitting

In this phase, the dataset is divided into two parts, namely training data and testing data, which aim to build and evaluate machine learning models systematically and increase performance. The proportion used is 20% of the data used for testing data and the remaining 80% used for training data, this is used to offer a good balance between the two, providing enough data for training while still leaving enough data for evaluation.

3.4. GNNs Model

GNNs are part of DL but can be used for classification tasks which means the dataset contains the final or target label class, this study uses a dataset containing classes. 2 main types of classification can be done, namely node classification which aims to predict the label of an individual node, and graph classification which aims to predict the label of the entire graph, which in essence this model is effective in handling classification tasks. GNNs are based on the concept of message passing and aggregation of neighbor information in a graph. There are 2 main components, namely Node Representations (Features) and Graph Convolution (Message Passing) and the basic formula:

$$h_v^{(l+1)} = \sigma\,(W^{(1)}.AGGREGATE\ (\left\{h_u^{(1)}: u \in N\ (v)\right\} +\ b^{(1)}) \tag{1}$$

where $h_v^{(l+1)}$ is a feature of node v on the layer $l+1$. $W^{(1)}$ is a trainable weight matrix on the layer l. AGGREGATE is an aggregation function that combines neighboring features$\in N\ (v)$. $b^{(1)}$ is bias on the layer $l\sigma$ is a non-linear activation function.

One metric in measuring model accuracy is the most commonly used in classification problems including GNNs. The accuracy formula for GNNs is the same as other ML models, which is the ratio of correct predictions (positive and negative) to the entire data, here is the formula:

$$Accuracy = \frac{(True\ Positif + True\ Negatif)}{(True\ Positif + False\ Positif + False\ Negatif + True\ Negatif)}\ x\ 100\% \tag{2}$$

$$Precision = \frac{True\ Positives}{True\ Positives + False\ Positives}\ x\ 100\% \tag{3}$$

$$Recall\ = \frac{True\ Positives}{True\ Positives + False\ Negatives}\ x\ 100\% \tag{4}$$

$$F1 - Score\ =\ 2\ x\ \frac{Recall * Precision}{Recall + Precision}\ x\ 100\% \tag{5}$$

Apart from accuracy, Precision is an important metric in measuring the model's ability to identify positive predictions that are truly positive. A high precision value indicates that this model is accurate in predictions. Next, Recall is an important metric in measuring the model's ability to identify all positive instances in the dataset correctly. A high recall value indicates that this model has good sensitivity in recognizing patterns and features, the formula is also general like other classification models. Finally, F1-Score is an important metric for evaluating performance which is capable of balancing precision and recall. With the highest F1-Score value, it shows very good positive performance and if there is an imbalance, it gives a more realistic picture of model performance than just using precision or recall alone. In this study, using GNNs, the accuracy results were seen in describing how good the model was in identifying or predicting the correct class of input data.

Accuracy, Precision, Recall, and F1-Score are critical in addressing the challenges of internet security prediction, especially regarding false positives and false negatives. Accuracy measures correct predictions but can be misleading in imbalanced data, where the model has high accuracy despite failing to detect a threat. Precision helps reduce false positives by ensuring that only legitimate threats are alerted, while Recall is important for reducing false negatives, ensuring that most of the real threats are detected. F1-Score provides a balance between precision and recall, ensuring that the model is effective in detecting threats while minimizing errors in both false positives and false negatives.

The architecture used in GNNs in this study is with Graph Convolutional Networks (GCN) because it is simpler, but more suitable for graphs that have a clear and consistent relationship structure. It is also very suitable for graph-based classification data because of its ability to capture and process relationships between nodes in the graph structure. This model can update node representations by combining information from node neighbors through convolution layers, allowing for a better understanding of patterns and relationships in graphs [114].

3.5. Feature Importance CatBoost Model

This model is well suited for finding the most influential features, especially in supervised learning or classification. Although there is no single mathematical formula as it involves several complex concepts, the Gradient Boosting approach builds predictions incrementally. One of the most common ways to select features is by using feature importance, a technique to measure how much each feature contributes to the prediction result. Several important metrics for assessing features, including Shape Values and Feature Importance based on the boosting algorithm, can help in identifying the features that have the most impact on performance. In general, the formula is as follows:

$$F_m(x) =\ F_{m-1}(x) +\ \gamma_m h_m(x) \tag{6}$$

where $F_m(x)$ is a model in iteration $m. h_m$ is a tree on iteration $m. \gamma_m$ is a weighting coefficient determined by minimizing the loss function. $F_{m-}(x)$ is a model in the previous iteration.

In this study, the highest or influential feature searcher with CatBoost is because it has advantages compared to other feature search models and is very effective and accurate in feature selection and seeing how much each feature contributes to model performance. Primarily, the use of the Catboost model provides a very good and easy-to-understand interpretability model from "black box" data.

3.6. End

Finally, which results are the highest from GNNs testing, and which features have the highest influence with CatBoost, with the results of these two tests there is potential for synergy and contributing to the creation of internet security prediction models.

4. RESULTS AND DISCUSSION

This section presents the results of an analysis of the performance of the GNNs and CatBoost algorithms. From GNNs, an evaluation was carried out by measuring accuracy, precision, sensitivity, and F1 score, looking at the extent to which the model can classify data and its ability to improve prediction results [115], [116].

Meanwhile, CatBoost focuses on identifying the most significant features that affect model predictions, providing insight into the main factors that contribute to the model's performance [117], [118]. This goal is to compare the effectiveness of the two algorithms in the context of testing different models, as well as to examine how each model handles the complexity and characteristics of the available data.

4.1. Testing With GNNs Model

Before testing, preprocessing is done by first handling missing values by replacing them with 0 using *df.fillna(0, inplace=True)*. Next, categorical features such as 'web_traffic', 'domain_age', 'domain_registration_length', and others are encoded using LabelEncoder to convert categories into numeric values. For numeric features, although not required, scaling can be done with StandardScaler to ensure the features are on a uniform scale, although this section is commented out in the code. Then, the data is prepared for PyTorch Geometric by creating edge_indexes that describe the relationship between nodes and preparing node (x) and label (y) features by converting the data into PyTorch tensor format. After that, the data is split into training and testing sets using train_test_split, followed by converting the training and testing data into PyTorch tensors for GNN model training.

As shown in in Table 3, it is found that the GNNs model achieves an accuracy of 93.58% which indicates that the model has very good performance and is accurate overall. The Precision score of 93.63% indicates that the model reliably identifies positive predictions with minimal false positives, making it the most reliable metric among the four in identifying positive classes and minimizing false positive prediction errors. The Recall results show the results of 93.58% of all positive cases, the value is the same as Accuracy which means it is not only accurate but effective in detecting positive classes. It can be seen that the precision and recall values are almost equal, indicating a balance in managing both false positives and false negatives. This also indicates that the system works effectively in both aspects, avoiding false positive predictions, while capturing almost all positive cases in the dataset. The F1-Score results which are a balance between Precision and Recall show the results of 93.55%, meaning it has a very good balance.

Table 3. GNNs Model Testing Results

Testing Results			
Accuracy	**Precision**	**Recall**	**F1-Score**
0.9358	0.9363	0.9358	0.9355

In the context of designing internet security prediction applications from matrix results, both Precision and Recall play a very important role, depending on the priority of the application objectives. If Precision is higher or more important, it avoids giving wrong warnings, and it tends to be more selective in threat prediction, but if Recall is higher or more important, it identifies all threats, but the risk could be the wrong threat. Ideally, one should look at the balance between precision and recall for optimal results.

From model testing with parameters covering various aspects that are important for improving model performance, such as the number of layers, representation dimension size, activation function, batch size, learning rate, and regularization techniques such as L2 regularization as adding penalties to large weights, encouraging small weights, and dropout which is a technique where some units are randomly removed during training.

As shown in Table 4, the dataset was tested with similar models, such as Graph Convolutional Networks (GCNs), Graph Attention Networks (GATs), and GraphSAGE, to compare the performance. Using these models, results show that while GNNs provide excellent results, other similar models can also provide competitive performance, depending on the graph structure and the type of relationship between nodes in the dataset. However, based on the comparison matrix, it proved to provide excellent performance compared to other models as well due to its ability to utilize the graph structure directly in the learning process, allowing it to capture complex relationship patterns between nodes that are difficult for other models to understand.

Table 4. Comparison of similar models with GNNs

Testing Results				
Model	**Accuracy**	**Precision**	**Recall**	**F1-Score**
GCNs	0.9303	0.9305	0.9303	0.9302
GATs	0.9303	0.9305	0.9303	0.9302
GraphSAGE	0.9358	0.9358	0.9358	0.9358

4.2. Testing With CatBoost Model

From the feature importance analysis in Fig. 3, it is evident that be seen the ranking or order of how much influence or relevance each variable has on the model prediction, especially displaying the visual or interpretability of each feature. From the test results with the CatBoost model, the highest features in internet security prediction are URL_of_Anchor, SSLfinal_State, and web_traffic. The URL_of_Anchor feature has the highest or most significant influence because there is information that reflects the relevance and context of a web page. Also, this feature has a lot of keywords or comes from trusted sources which can increase the authority and relevance of the page, thus playing a role in determining the prediction results. Thus, the characteristics of URL_of_Anchor are important indicators in evaluating the quality and security of the analyzed website. In addition, the SSLfinal_State and web_traffic features show a very large influence. The SSLfinal_State feature reflects the security of the website, with sites that use valid and well-configured SSL/TLS encryption often considered more trustworthy by users and systems. This highlights the importance of the security aspect in model assessment. On the other hand, web_traffic measures how many and how often users visit the site, the higher the visits, the more correlated with the quality and relevance of the site, making it an important factor in the analysis. Overall, these three features provide deep insights into the quality and trustworthiness of the website and contribute significantly to the predictions generated by the CatBoost model.

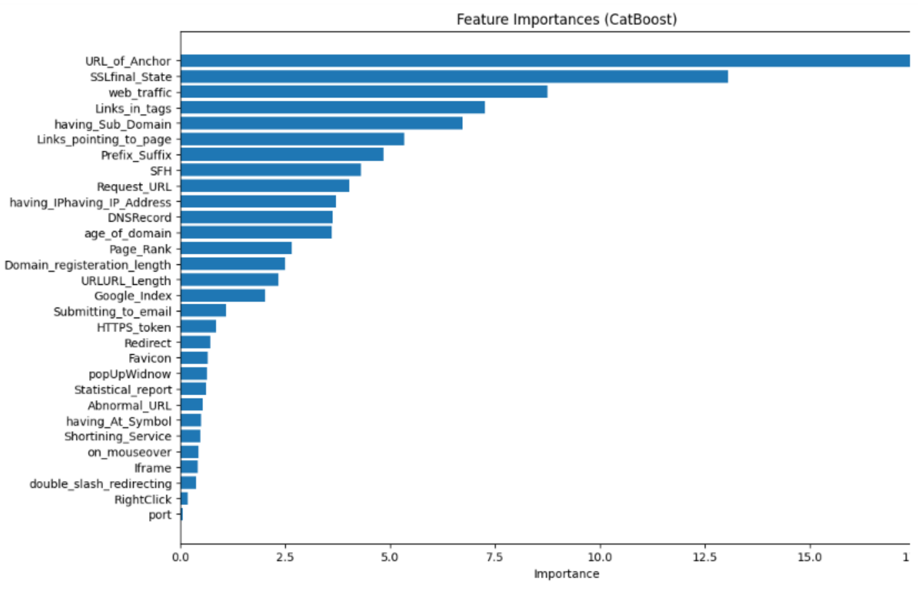

Fig. 3. Feature Importances (CatBoost)

URL_of_Anchor, SSLfinal_State, and web_traffic features can help security practitioners and web developers increase website trust and security. URL_of_Anchor provides knowledge about the reputation of the source that leads to a secure site and, ensures it comes from a trusted site. SSLfinal_State ensures the use of secure encryption (HTTPS) to protect user data increase visitor trust, and can improve a site's SEO ranking. Meanwhile, web traffic can be used to detect potential threats such as DDoS attacks through unnatural traffic spikes, and help developers improve site performance and security based on visitor behavior patterns. These three features can help design an early warning system that enables rapid detection and response to threats in internet security.

As shown in Table 5, the CatBoost model achieves an Accuracy of 96.15%, Precision of 95.70%, Recall of 97.60%, and F1-Score of 96.64%, which shows excellent and balanced performance in all evaluation metrics. The high AUC-ROC (99.44%) confirms the model's ability to distinguish between positive and negative classes. The almost equal values of Accuracy, Precision, Recall, and F1-Score indicate that the model has a consistent and effective performance, which is likely due to the balanced class distribution in the dataset. This successfully minimizes errors on both sides (false positives and false negatives), resulting in accurate and balanced predictions. Based on the matrix testing results, the CatBoost model shows optimal performance in recognizing and classifying both positive and negative classes, with very low false positive and false negative rates, suggesting that it is highly reliable in applications that require high accuracy and sensitive detection.

Table 5. CatBoost Model Testing Results

Testing Results				
Accuracy	**Precision**	**Recall**	**F1-Score**	**AUC-ROC**
0.9615	0.9570	0.9760	0.9664	0.9944

This research has limitations, especially the difficulty of generalizing to new data sets or domains, which can lead to performance degradation if the model is not flexible enough to handle different data variations. Overfitting is a risk when the metric is high on the training data but the model fails on the test data, due to overfitting to the specific pattern of the training data. In addition, poor data quality or bias in the training data can affect model results, while inappropriate metric selection can give a false picture of model performance, especially in the context of class-imbalanced data. Highly complex models are also prone to performance degradation when applied to different scenarios, and in the context of security, models can struggle to detect new threats that are not present in the training data, requiring better adaptation to evolving threats. If there is a trade-off between the two models then it is possible, that GNNs predictions are very good and graph analysis is very powerful but interpretation is very difficult and poor. CatBoost is very good easy to understand in interpretability, and very suitable for decision making and modeling predictions.

5. CONCLUSION AND FUTURE WORK

The test results of the Graph Neural Networks (GNNs) and CatBoost models provide different but complementary insights into performance and important features in data analysis. The GNNs model shows excellent performance with an Accuracy of 93.58%, Precision of 93.63%, Recall of 93.58%, and F1-Score of 93.55%. This shows that the GNNs model can provide very accurate and balanced predictions in terms of Precision and Recall, demonstrating its ability to understand and generalize complex patterns in structured data such as graphs.

The CatBoost model test results identified important features that influence the prediction of URL_of_Anchor, SSLfinal_State, and web_traffic. Of the three features, it was proven to be significant in determining the quality and credibility of the website, with URL_of_Anchor focusing on the relevance of the link to be visited, SSLfinal_State assessing the security aspect, and web_traffic reflecting the popularity of the many visits.

This research highlights the utility of combining Graph Neural Networks (GNNs) for pattern recognition and CatBoost for feature interpretation, providing a dual perspective that improves internet security predictions. The combination of insights from these two models not only enables accurate and balanced predictions thanks to GNNs' ability to leverage graph data structures but also provides a clearer understanding of the analyzed data through the interpretation of features generated by CatBoost. By identifying key features that drive predictions, CatBoost offers actionable insights to improve web security, while GNNs provide a solid foundation for decision-making by leveraging relationships between entities in the graph. This combination provides a more comprehensive and in-depth understanding of the data, which is crucial in the context of strengthening internet security systems.

It is expected that in the future we can focus on exploring and integrating new features to improve the accuracy and capability of the model in predicting internet security, in addition to more features and data to capture wider variability in internet security patterns. Moreover, we can test with other DL models such as CNN, RNN, and others or compare additionally, employing interpretability tools like LIME and SHAP can enhance feature understanding, while deploying the models in adversarial environments can validate their robustness against sophisticated cyber threats.

REFERENCES

[1] T. T. Kwon *et al*, "How to decentralize the internet: A focus on data consolidation and user privacy," *Computer Networks*, vol. 234, p. 109911, 2023, https://doi.org/10.1016/j.comnet.2023.109911.

[2] A. Szymkowiak, *et al.*, "Information technology and Gen Z: The role of teachers, the internet, and technology in the education of young people," *Technology in Society*, vol. 65, p. 101565, 2021, https://doi.org/10.1016/j.techsoc.2021.101565.

[3] A. Roukounaki *et al*, "Scalable and Configurable End-to-End Collection and Analysis of IoT Security Data: Towards End-to-End Security in IoT Systems," *Global IoT Summit (GIoTS)*, pp. 1-6, 2019, https://doi.org/10.1109/GIOTS.2019.8766407.

[4] M. Alazab, S. Hong, and J. Ng, "Louder bark with no bite: Privacy protection through the regulation of mandatory data breach notification in Australia," *Future Generation Computer Systems*, vol. 116, pp. 22-29, 2021, https://doi.org/10.1016/j.future.2020.10.017.

[5] W. Li, and Z. Yang, "Landscape design of urban culture transmission based on the regional information security of Internet of Things," *Heliyon*, vol. 10, no. 15, p. e35042, 2024, https://doi.org/10.1016/j.heliyon.2024.e35042.

[6] J. B. B. Pea-Assounga *et al*, "Effect of financial innovation and stakeholders' satisfaction on investment decisions: does internet security matter?," *Heliyon*, vol, 10, no. 6, p. e27242, 2024, https://doi.org/10.1016/j.heliyon.2024.e27242.
[7] L. Tawalbeh *et al*, "IoT Privacy and Security: Challenges and Solutions," *Applied Sciences*. vol. 10, no. 12, p. 4102, 2020; https://doi.org/10.3390/app10124102.
[8] P. R. Kanna, and P. Santhi, "Exploring the landscape of network security: a comparative analysis of attack detection strategies," *J Ambient Intell Human Comput,* vol. 15, pp. 3211–3228, 2024, https://doi.org/10.1007/s12652-024-04794-y.
[9] S. J. Holmen, "Situational Crime Prevention, Advice Giving, and Victim-Blaming," *Philosophia,* vol. 52, pp. 325–340, 2024, https://doi.org/10.1007/s11406-024-00729-1.
[10] I. Chenchev, "Framework for Multi-factor Authentication with Dynamically Generated Passwords," *Advances in Information and Communication. FICC*, *Lecture Notes in Networks and Systems*, vol 652, 2023, https://doi.org/10.1007/978-3-031-28073-3_39.
[11] A. Girma, M. A. Guo, and J. Irungu, "Identifying Shared Security Vulnerabilities and Mitigation Strategies at the Intersection of Application Programming Interfaces (APIs), Application-Level and Operating System (OS) of Mobile Devices," *Proceedings of the Future Technologies Conference (FTC)*, *Lecture Notes in Networks and Systems*, vol 560, 2022, https://doi.org/10.1007/978-3-031-18458-1_34.
[12] A. M. Sakshi, and A. K, Sharma, "A survey on blockchain based IoT forensic evidence preservation: research trends and current challenges," *Multimed Tools Appl*, vol. 83, pp. 42413–42458, 2024, https://doi.org/10.1007/s11042-023-17104-z.
[13] P. Victor *et al*, "IoT malware: An attribute-based taxonomy, detection mechanisms and challenges," *Peer-to-Peer Netw. Appl*. vol. 16, pp. 1380–1431, 2023, https://doi.org/10.1007/s12083-023-01478-w.
[14] S.Rudrakar, and P, Rughani, "IoT based Agriculture (Ag-IoT): A detailed study on Architecture, Security and Forensics, Information," *Processing in Agriculture*, 2023, https://doi.org/10.1016/j.inpa.2023.09.002.
[15] R. Kumar *et al,* "Machine and deep learning methods for concrete strength Prediction: A bibliometric and content analysis review of research trends and future directions," *Applied Soft Computing*, vol. 164, p. 111956, 2024, https://doi.org/10.1016/j.asoc.2024.111956.
[16] J. Sun *et al*, "Hybrid deep learning approach for rock tunnel deformation prediction based on spatio-temporal patterns," *Underground Space*, vol. 20, pp. 100-118, 2024, https://doi.org/10.1016/j.undsp.2024.04.008.
[17] F. Alhaek *et al*, "Learning spatial patterns and temporal dependencies for traffic accident severity prediction: A deep learning approach," *Knowledge-Based Systems*, vol. 286, p. 111406, 2024, https://doi.org/10.1016/j.knosys.2024.111406.
[18] G. Zare, N. J. Navimipour, M. Hosseinzadeh, and A. Sahafi, "Network link prediction via deep learning method: A comparative analysis with traditional methods," *Engineering Science and Technology, an International Journal*, vol. 56, p. 101782, 2024, https://doi.org/10.1016/j.jestch.2024.101782.
[19] D. V. Nguyen, Y. Choo, and D. Kim, "Deep learning application for nonlinear seismic ground response prediction based on centrifuge test and numerical analysis," *Soil Dynamics and Earthquake Engineering*, vol. 182, p. 108733, 2024, https://doi.org/10.1016/j.soildyn.2024.108733.
[20] H. Jebnoun *et al*, "Clones in deep learning code: what, where, and why?," *Empir Software Eng,* vol. 27, no. 4, p. 84, 2022. https://doi.org/10.1007/s10664-021-10099-x.
[21] L, Wang, Z. Zhu, dan X, Zhao, "Dynamic predictive maintenance strategy for system remaining useful life prediction via deep learning ensemble method," *Reliability Engineering & System Safety*, vol. 245, p. 110012, 2024, https://doi.org/10.1016/j.ress.2024.110012.
[22] G. E. Vadivu, and T. Muthusamy, "Synthesis of deep learning technique for social distance monitoring in pandemic areas," *Multimed Tools Appl,* vol. 83, pp. 30361–30376, 2024, https://doi.org/10.1007/s11042-023-16681-3.
[23] U. H. Atasever, and E, Tercan, "Deep learning-based burned forest areas mapping via Sentinel-2 imagery: a comparative study*," Environ Sci Pollut Res*, vol. 31, pp. 5304–5318, 2024, https://doi.org/10.1007/s11356-023-31575-5.
[24] S. Yang *et al*, "Improving Mapping Accuracy of Smallholder Potato Planting Areas by Embedding Prior Knowledge into a Novel Multi-temporal Deep Learning Network" *Potato Res*, pp. 1-31, 2024, https://doi.org/10.1007/s11540-024-09769-2.
[25] K. Sharma, G. K. Sethi, and R. K. Bawa, "A comparative analysis of deep learning and deep transfer learning approaches for identification of rice varieties." *Multimed Tools Appl*, pp. 1-18, 2024, https://doi.org/10.1007/s11042-024-19126-7.
[26] T. Wang *et al*, "COFNet: A deep learning model to predict the specific surface area of covalent-organic frameworks using structural images and statistic features," *Chemical Physics Letters*, vol. 847, p. 141383, 2024, https://doi.org/10.1016/j.cplett.2024.141383.
[27] Qi Liao *et al*, "Probing the capacity of a spatiotemporal deep learning model for short-term PM2.5 forecasts in a coastal urban area," *Science of The Total Environment*, vol, 950, p. 175233, 2024, https://doi.org/10.1016/j.scitotenv.2024.175233\.
[28] R. E. Nogales, and M. E. Benalcázar, "Analysis and Evaluation of Feature Selection and Feature Extraction Methods," *Int J Comput Intell Syst*, vol. 16, p. 153, 2023, https://doi.org/10.1007/s44196-023-00319-1.
[29] B. Beceiro *et al*, "CUDA acceleration of MI-based feature selection methods," *Journal of Parallel and Distributed Computing*, vol. 190, p. 104901, 2024, https://doi.org/10.1016/j.jpdc.2024.104901.

[30] K. Zhang *et al*, "Enhancing IoT (Internet of Things) feature selection: A two-stage approach via an improved whale optimization algorithm," *Expert Systems with Applications*, vol. 256, p. 124936, 2024, https://doi.org/10.1016/j.eswa.2024.124936.
[31] A. Moslemi, and M. Bidar, "Dual-dual subspace learning with low-rank consideration for feature selection," *Physica A: Statistical Mechanics and its Applications*, vol. 651, p. 129997, 2024, https://doi.org/10.1016/j.physa.2024.129997.
[32] K. Okoye, and S. Hosseini, "Correlation Tests in R: Pearson Cor, Kendall's Tau, and Spearman's Rho," I*n: R Programming. Springer*, pp. 247-277, 2024, https://doi.org/10.1007/978-981-97-3385-9_12.
[33] P. Lin *et al,* "An Intelligent Depth Correction Method for Logging Curves Based on Pearson Correlation Coefficient and DTW," *Proceedings of the International Field Exploration and Development Conference*, pp. 102-114, 2023, https://doi.org/10.1007/978-981-97-0479-8_8.
[34] R. Okunev, "Pearson Correlation and Using the Excel Linear Trend Equation and Excel Regression Output," *In: Analytics for Retail*, pp. 83-106, 2022, https://doi.org/10.1007/978-1-4842-7830-7_8.
[35] B.F. Darst, K. C. Malecki, and C. D. Engelman, "Using recursive feature elimination in random forest to account for correlated variables in high dimensional data," *BMC Genet*, vol. 19, pp. 1-6, 2018, https://doi.org/10.1186/s12863-018-0633-8.
[36] N. R. Abid-Althaqafi, and H. A. Alsalamah, "The Effect of Feature Selection on the Accuracy of X-Platform User Credibility Detection with Supervised Machine Learning," *Electronics*, vol. 13, no. 1, p. 20, 2024, https://doi.org/10.3390/electronics13010205.
[37] C. Fraser, "Association Between Categorical Variables: Contingency Analysis with Chi Square," *In: Business Statistics for Competitive Advantage with Excel and JMP*, 2024, https://doi.org/10.1007/978-3-031-42555-4_3
[38] B. K. Das *et al*, "Square Test of Significance," *In: Concept Building in Fisheries Data Analysis*, *Springer*, pp. 81-942022, 2022, https://doi.org/10.1007/978-981-19-4411-6_5.
[39] D. Al-Shammary *et al*, "Efficient ECG classification based on Chi-square distance for arrhythmia detection," *Journal of Electronic Science and Technology*, vol. 22, no. 2, p. 100249, 2024, https://doi.org/10.1016/j.jnlest.2024.100249.
[40] C. Yang *et al*, "How can SHAP (SHapley Additive exPlanations) interpretations improve deep learning based urban cellular automata model?," *Computers, Environment and Urban Systems*, vol. 111, p. 102133, 2024, https://doi.org/10.1016/j.compenvurbsys.2024.102133.
[41] A. S. Antonini *et al*, "Machine Learning model interpretability using SHAP values: Application to Igneous Rock Classification task," *Applied Computing and Geosciences*, vol. 23, p. 100178, 2024, https://doi.org/10.1016/j.acags.2024.100178.
[42] X. Cheng *et al*, "Predicting response to CCRT for esophageal squamous carcinoma by a radiomics-clinical SHAP model," *BMC Med Imaging,* vol. 23, no. 1, p. 145, 2023, https://doi.org/10.1186/s12880-023-01089-0.
[43] J. Shin, "Feasibility of local interpretable model-agnostic explanations (LIME) algorithm as an effective and interpretable feature selection method: comparative fNIRS study," *Biomed. Eng. Lett.* vol. 13, pp. 689–703, 2023, https://doi.org/10.1007/s13534-023-00291-x.
[44] C. Hsu *et al*, "Artificial Intelligence Model Interpreting Tools: SHAP, LIME, and Anchor Implementation in CNN Model for Hand Gestures Recognition," *In: Technologies and Applications of Artificial Intelligence. Communications in Computer and Information Science,* vol. 2074, 2023, https://doi.org/10.1007/978-981-97-1711-8_2.
[45] T. V. Krishnamoorthy *et al*, "A novel NASNet model with LIME explanability for lung disease classification," *Biomedical Signal Processing and Control*, vol. 93, pp. 106114, 2024, https://doi.org/10.1016/j.bspc.2024.106114.
[46] G. Manikandan *et al*, "Classification models combined with Boruta feature selection for heart disease prediction," *Informatics in Medicine Unlocked*, vol. 44, p. 101442, 2024, https://doi.org/10.1016/j.imu.2023.101442.
[47] H. Luo *et al*, "SHAP based predictive modeling for year all-cause readmission risk in elderly heart failure patients: feature selection and model interpretation," *Sci Rep*, vol. 14, p. 17728, 2024, https://doi.org/10.1038/s41598-024-67844-7.
[48] F. Türk, "Investigation of machine learning algorithms on heart disease through dominant feature detection and feature selection," *SIViP,* vol. 18, pp. 3943–3955, 2024, https://doi.org/10.1007/s11760-024-03060-0.
[49] A. Dardzińska-Głębocka, and M. Zdrodowska, "Analysis children with disabilities self-care problems based on selected data mining techniques," *Procedia Computer Science*, vol. 192, pp. 2854-2862, 2021, https://doi.org/10.1016/j.procs.2021.09.056.
[50] Md. S. H. Shaon *et al*, "A comparative study of machine learning models with LASSO and SHAP feature selection for breast cancer prediction," *Healthcare Analytics*, vol. 6, p. 100353, 2024, https://doi.org/10.1016/j.health.2024.100353.
[51] H. Chereda, A. Leha, and T. Beibarth, "Stable feature selection utilizing Graph Convolutional Neural Network and Layer-wise Relevance Propagation for biomarker discovery in breast cancer," *Artificial Intelligence in Medicine*, vol. 151, p. 102840, 2024, https://doi.org/10.1016/j.artmed.2024.102840.
[52] A. Et-touri *et al*, "Comparison of Feature Selection Methods for Breast Cancer Prediction," *International Conference on Advanced Intelligent Systems for Sustainable Development (AI2SD'2023)*. pp. 272-282, 2023, 2023, https://doi.org/10.1007/978-3-031-54318-0_23.
[53] S. P. Jakhar *et al*, "Brain tumor detection with multi-scale fractal feature network and fractal residual learning," *Applied Soft Computing*, vol. 153, p. 111284, 2024, https://doi.org/10.1016/j.asoc.2024.111284.

[54] X. Liu *et al*, "A hierarchical attention-based feature selection and fusion method for credit risk assessment," *Future Generation Computer Systems,* vol. 160, pp. 537-546, 2024, https://doi.org/10.1016/j.future.2024.06.036.
[55] Y. Zhao *et al*, "Carbon futures price forecasting based on feature selection," *Engineering Applications of Artificial Intelligence*, vol. 135, p. 108646, 2024, https://doi.org/10.1016/j.engappai.2024.108646.
[56] J. Wang, and Y. Dong, "An interpretable deep learning multi-dimensional integration framework for exchange rate forecasting based on deep and shallow feature selection and snapshot ensemble technology," *Engineering Applications of Artificial Intelligence*, vol. 133, Part C, p. 108282, 2024, https://doi.org/10.1016/j.engappai.2024.108282.
[57] H. Eskandari *et al*, "Innovative framework for accurate and transparent forecasting of energy consumption: A fusion of feature selection and interpretable machine learning," *Applied Energy*, vol. 366, p. 123314, 2024, https://doi.org/10.1016/j.apenergy.2024.123314.
[58] Q. Qiao *et al*, "An interpretable multi-stage forecasting framework for energy consumption and CO2 emissions for the transportation sector," *Energy*, vol. 286, p. 129499, 2024, https://doi.org/10.1016/j.energy.2023.129499.
[59] M. Sharma *et al*, "Ensemble learning for prominent feature selection and electric power prediction in agriculture sector," *Multimed Tools Appl*, pp. 1-28, 2024, https://doi.org/10.1007/s11042-024-18179-y.
[60] M. Radwan *et al*, "Potato Leaf Disease Classification Using Optimized Machine Learning Models and Feature Selection Technique," *Potato Res*, pp. 1-25, 2024, https://doi.org/10.1007/s11540-024-09763-8
[61] W. Cao *et al*, "A STAM-LSTM model for wind power prediction with feature selection," *Energy*, vol. 296, p. 131030, 2024, https://doi.org/10.1016/j.energy.2024.131030.
[62] G. Nasreen *et al*, "Email spam detection by deep learning models using novel feature selection technique and BERT," *Egyptian Informatics Journal*, vol. 26, p. 100473, 2024, https://doi.org/10.1016/j.eij.2024.100473.
[63] G. Kapoor, and N. Wichitaksorn, "Electricity price forecasting in New Zealand: A comparative analysis of statistical and machine learning models with feature selection," *Applied Energy,* vol. 347, p. 121446, 2023, https://doi.org/10.1016/j.apenergy.2023.121446.
[64] D. Sagar, and M. Saidireddy, "Security Measurement in LTE/LTE-A Network Based on zS-LR Feature Selection Technique and UM-tGAN Attack Detection Technique," *Expert Systems with Applications*, vol. 231, p. 120703, 2023, https://doi.org/10.1016/j.eswa.2023.120703.
[65] R. Yadav, I. Sreedevi, and D. Gupta, "Augmentation in performance and security of WSNs for IoT applications using feature selection and classification techniques," *Alexandria Engineering Journal*, vol. 65, pp. 461-473, 2023, https://doi.org/10.1016/j.aej.2022.10.033.
[66] Q. B. Baker, and A. Samarneh, "Feature selection for IoT botnet detection using equilibrium and Battle Royale Optimization," *Computers & Security*, vol. 147, p. 104060, 2024, https://doi.org/10.1016/j.cose.2024.104060.
[67] V. Roblekasas *et al*, "The Interaction between Internet, Sustainable Development, and Emergence of Society 5.0," *Data*, vol. 5, p. 80, 2020, https://doi.org/10.3390/data5030080.
[68] R. Mohan, "The effect of population growth, the pattern of demand and of technology on the process of urbanization," *Journal of Urban Economics*, vol. 15, no. 2, pp. 125-156, 1984, https://doi.org/10.1016/0094-1190(84)90011-1.
[69] M. Lubis, and D. O. D. Handayani, "The relationship of personal data protection towards internet addiction: Cybercrimes, pornography and reduced physical activity," *Procedia Computer Science*, vol. 197, pp.151-161, 2022, https://doi.org/10.1016/j.procs.2021.12.129.
[70] R. Ayachi, Y. Said, and A. B. Abdellali, "Pedestrian Detection Based on Light-Weighted Separable Convolution for Advanced Driver Assistance Systems," *Neural Process Lett*, vol. 52, pp. 2655–2668, 2020, https://doi.org/10.1007/s11063-020-10367-9.
[71] M. Afif e*t al*, "A Transfer Learning Approach for Indoor Object Identification," *SN COMPUT. SCI,* vol. 2, p. 424, 2021, https://doi.org/10.1007/s42979-021-00790-7.
[72] R. Ayachi e*t al*, "Traffic Signs Detection for Real-World Application of an Advanced Driving Assisting System Using Deep Learning," *Neural Process Lett,* vol. 51, pp. 837–851, 2020, https://doi.org/10.1007/s11063-019-10115-8.
[73] Y. Said e*t al*, "Medical Images Segmentation for Lung Cancer Diagnosis Based on Deep Learning Architectures," *Diagnostics*. vol. 13, no. 3, p. 546, 2023, https://doi.org/10.3390/diagnostics13030546.
[74] F. Mohammad, S. Al-Ahmadi, and J. Al-Muhtadi, "Deep Learning Based Cyber Event Detection from Open-Source Re-Emerging Social Data," *Computers, Materials and Continua*, vol. 76, no. 2, pp. 1423-1438, 2023, https://doi.org/10.32604/cmc.2023.035741.
[75] M. Alshehri e*t al*, "Character-level word encoding deep learning model for combating cyber threats in phishing URL detection," *Computers and Electrical Engineering*, vol. 100, p. 107868, 2022, https://doi.org/10.1016/j.compeleceng.2022.107868.
[76] D. Chen, P. Wawrzynski, and Z. Lv, "Cyber security in smart cities: A review of deep learning-based applications and case studies," *Sustainable Cities and Society*, vol. 66, p. 102655, 2021, https://doi.org/10.1016/j.scs.2020.102655.
[77] Y. Sun e*t al*, "GTC: GNN-Transformer co-contrastive learning for self-supervised heterogeneous graph representation," *Neural Networks*, vol. 181, p. 106645, 2024, https://doi.org/10.1016/j.neunet.2024.106645.
[78] H. A. Mohamed e*t al*, "Locality-aware subgraphs for inductive link prediction in knowledge graphs," *Pattern Recognition Letters*, vol. 167, pp. 90-97, 2023, https://doi.org/10.1016/j.patrec.2023.02.004.
[79] X. Li e*t al*, "Table Structure Recognition and Form Parsing by End-to-End Object Detection and Relation Parsing, Pattern Recognition," vol. 132, p. 108946, 2022, https://doi.org/10.1016/j.patcog.2022.108946.

[80] N. Das et *al*, "Integrating sentiment analysis with graph neural networks for enhanced stock prediction: A comprehensive survey," *Decision Analytics Journal*, vol. 10, p. 100417, 2024, https://doi.org/10.1016/j.dajour.2024.100417.
[81] I. D. Mienye, T. G. Swart, and G. Obaido, "Recurrent Neural Networks: A Comprehensive Review of Architectures, Variants, and Applications," *Information*, vol. 15, no. 9, p. 517, 2024, https://doi.org/10.3390/info15090517.
[82] L. Alzubaidi *et al,* "Review of deep learning: concepts, CNN architectures, challenges, applications, future directions," *J Big Data*, vol. 8, p. 53, 2021, https://doi.org/10.1186/s40537-021-00444-8.
[83] R. Zhao e*t al*, "A two-stage CFD-GNN approach for efficient steady-state prediction of urban airflow and airborne contaminant dispersion," *Sustainable Cities and Society*, vol. 112, p. 105607, 2024, https://doi.org/10.1016/j.scs.2024.105607.
[84] A. A. Makhdnomi, and I. A. Gillani, "GNN-based passenger request prediction," *Transportation Letters*, pp. 1-15, 2024, https://doi.org/10.1080/19427867.2023.2283949.
[85] Y. Lei e*t al*, "GNN-fused CapsNet with multi-head prediction for diabetic retinopathy grading," *Engineering Applications of Artificial Intelligence*, vol. 133, Part A, p. 107994, 2024, https://doi.org/10.1016/j.engappai.2024.107994.
[86] M. Farreras e*t al*, "Improving Network Delay Predictions Using GNNs," *J Netw Syst Manage*, vol. 31, p. 65, 2023, https://doi.org/10.1007/s10922-023-09758-9.
[87] M. Davidson, and D. Moodley, "ST-GNNs for Weather Prediction in South Africa," *In Southern African Conference for Artificial Intelligence Research*, pp. 93-107, 2022, https://doi.org/10.1007/978-3-031-22321-1_7
[88] N. Q. K. Le, "Predicting emerging drug interactions using GNNs," *Nat Comput Sci*, vol. 3, pp. 1007–1008, 2023, https://doi.org/10.1038/s43588-023-00555-7.
[89] Q, Dang, "Detecting Obfuscated Malware Using Graph Neural Networks," *Power Engineering and Intelligent Systems. PEIS, 2023. Lecture Notes in Electrical Engineering*, pp. 15-25, 2023, https://doi.org/10.1007/978-981-99-7216-6_2.
[90] M. M. El-Gayar e*t al*, "A novel approach for detecting deep fake videos using graph neural network," *J Big Data*, vol. 11, p. 22, 2024, https://doi.org/10.1186/s40537-024-00884-y.
[91] M. Belaoued e*t al*, "Deep Learning for Windows Malware Analysis," *Cyber Malware. Security Informatics and Law Enforcement*, pp. 119-164, 2024, https://doi.org/10.1007/978-3-031-34969-0_6.
[92] A. Ghaffari e*t al*, "Securing internet of things using machine and deep learning methods: a survey." *Cluster Comput*, pp. 1-25, 2024, https://doi.org/10.1007/s10586-024-04509-0.
[93] S. C. Chelgani e*t al*, "CatBoost-SHAP for modeling industrial operational flotation variables – A "conscious lab" approach," *Minerals Engineering*, vol. 213, p. 108754, 2024, https://doi.org/10.1016/j.mineng.2024.108754.
[94] X. Feng, J. He, and B. Lu, "Accurate and generalizable soil liquefaction prediction model based on the CatBoost algorithm," *Acta Geophys*, vol. 72, pp. 3417–3426, 2024, https://doi.org/10.1007/s11600-024-01381-9.
[95] R. Taherdangkoo e*t al*, "Modeling unsaturated hydraulic conductivity of compacted bentonite using a constrained CatBoost with bootstrap analysis," *Applied Clay Science*, vol. 260, p. 107530, 2024, https://doi.org/10.1016/j.clay.2024.107530.
[96] J. T. Hancock, and T. M. Khoshgoftaar, "CatBoost for big data: an interdisciplinary review," *J Big Data,* vol. 7, p. 94, 2020, https://doi.org/10.1186/s40537-020-00369-8.
[97] A. A. Ibrahim e*t al*, "Comparison of the CatBoost Classifier with other Machine Learning Methods," *International Journal of Advanced Computer Science and Applications (IJACSA)*, vol. 11, no. 11, 2020, http://dx.doi.org/10.14569/IJACSA.2020.0111190.
[98] H. Qiu e*t al*, "Prediction of hydrogen storage in metal-organic frameworks using CatBoost-based approach," *International Journal of Hydrogen Energy*, vol. 79, pp. 952-961, 2024, https://doi.org/10.1016/j.ijhydene.2024.07.078.
[99] Y. Zhou e*t al*, "Remaining useful life prediction and state of health diagnosis of lithium-ion batteries with multiscale health features based on optimized CatBoost algorithm," *Energy*, vol. 300, p. 131575, 2024, https://doi.org/10.1016/j.energy.2024.131575.
[100] X. Wei e*t al*, "Risk assessment of cardiovascular disease based on SOLSSA-CatBoost model," *Expert Systems with Applications*, vol. 219, p. 119648, 2023, https://doi.org/10.1016/j.eswa.2023.119648.
[101] B. Dhananjay, and J. Sivaraman, "Analysis and classification of heart rate using CatBoost feature ranking model," *Biomedical Signal Processing and Control*, vol. 68, p. 102610, 2021, https://doi.org/10.1016/j.bspc.2021.102610.
[102] S. Aziz e*t al*, "A Framework for Cardiac Arrest Prediction via Application of Ensemble Learning Using Boosting Algorithms," *Procedia Computer Science,* vol. 235, pp. 3293-3304, 2024, https://doi.org/10.1016/j.procs.2024.04.311.
[103] Ye. Shiren e*t al*, "Interpretable prediction model for assessing diabetes complication risks in Chinese sufferers," *Diabetes Research and Clinical Practice*, vol. 209, p. 111560, 2024, https://doi.org/10.1016/j.diabres.2024.111560.
[104] H. F. Harumy, S. M. Hardi, and M. F. Al Banna, "EarlyStage Diabetes Risk Detection Using Comparison of Xgboost, Lightgbm, and Catboost Algorithms," *In International Conference on Advanced Information Networking and Applications,* pp. 12-24, 2024, https://doi.org/10.1007/978-3-031-57931-8_2.
[105] S. K. S. Modak, and V. K. Jha, "Diabetes prediction model using machine learning techniques," *Multimed Tools Appl,* vol. 83, pp. 38523–38549, 2024, https://doi.org/10.1007/s11042-023-16745-4.

[106] J. Dhar, and S. Roy, "Identification and diagnosis of cervical cancer using a hybrid feature selection approach with the bayesian optimization-based optimized catboost classification algorithm," *J Ambient Intell Human Comput*, vol. 15, pp. 3459–3477, 2024, https://doi.org/10.1007/s12652-024-04825-8.
[107] P. B. Dash e*t al*, "Efficient Ensemble Learning Based CatBoost Approach for Early-Stage Stroke Risk Prediction," *In Ambient Intelligence in Health Care: Proceedings of ICAIHC 2022*, pp. 475-483, 2022, https://doi.org/10.1007/978-981-19-6068-0_46.
[108] H. Lu, and X. Hu, "Enhancing Financial Risk Prediction for Listed Companies: A Catboost-Based Ensemble Learning Approach," *J Knowl Econ*, vol. 15, pp. 9824–9840, 2024, https://doi.org/10.1007/s13132-023-01601-5.
[109] B. Yu e*t al*, "Risk Assessment of Multi-Hazards in Hangzhou: A Socioeconomic and Risk Mapping Approach Using the CatBoost-SHAP Model," *Int J Disaster Risk Sci*, vol. 15, no. 4, pp. 640-656, 2024, https://doi.org/10.1007/s13753-024-00578-2.
[110] X. Wei e*t al*, "Evaluating ensemble learning techniques for stock index trend prediction: a case of China," *Port Econ J*, vol. 23, pp. 505–530, 2024, https://doi.org/10.1007/s10258-023-00246-1.
[111] S. Porkodi, and D. Kesavaraja, "Scammer identification using CatBoost in smart contract for enhancing security in blockchain network," *Wireless Netw*, vol. 30, pp. 1165–1186, 2024, https://doi.org/10.1007/s11276-023-03552-w.
[112] M. Aguga *et al,* "Detection of Phishing Websites from URLs Using Hybrid Ensemble-Based Machine Learning Technique," *In International Conference on Soft Computing and Data Mining,* pp. 11-22, 2024, https://doi.org/10.1007/978-3-031-66965-1_2.
[113] L. C. M. Liaw e*t al*, "A histogram SMOTE-based sampling algorithm with incremental learning for imbalanced data classification," *Information Sciences*, vol. 686, pp. 121193, 2025, https://doi.org/10.1016/j.ins.2024.121193.
[114] S. Zhang, H. Tong, J. Xu, and R. Maciejewki, "Graph convolutional networks: a comprehensive review," *Comput Soc Netw*, vol. 6, p. 11, 2019, https://doi.org/10.1186/s40649-019-0069-y.
[115] F. Damoun, H. Seba, and R. State, "Privacy-Preserving Behavioral Anomaly Detection in Dynamic Graphs for Card Transactions." *In International Conference on Web Information Systems Engineering*, pp. 286-301, 2024, https://doi.org/10.1007/978-981-96-0576-7_22.
[116] E. A. V. Fabiano, and M. Recamonde-Mendoza, "Prediction of Cancer-Related miRNA Targets Using an Integrative Heterogeneous Graph Neural Network-Based Method." *In Brazilian Conference on Intelligent Systems,* pp. 346-360, 2023, https://doi.org/10.1007/978-3-031-45392-2_23.
[117] H. Wang *et al,* "Feature selection strategies: a comparative analysis of SHAP-value and importance-based methods. *J Big Data*, vol. 11, p. 44, 2024, https://doi.org/10.1186/s40537-024-00905-w.
[118] A. M. Alsaffar, M. Nouri-Baygi, and H. M. Zolbanin, "Shielding networks: enhancing intrusion detection with hybrid feature selection and stack ensemble learning." *J Big Data*, vol. 11, p. 133, 2024, https://doi.org/10.1186/s40537-024-00994-7.

BIOGRAPHY OF AUTHORS

Aswan Supriyadi Sunge, is a computer science lecturer at Pelita Bangsa University, West Java, Indonesia. I completed a bachelor's degree in economics at Gorontalo University in 2004. M.Com degree in Computer Science at Nusa Mandiri University, Jakarta, Indonesia in 2013. He completed a Doctorate in Computer Science at Bina Nusantara University, Indonesia in 2023. His research interests are Artificial Intelligence, Machine Learning, Deep Learning, Data Mining, Text Mining, Image Processing, Clustering, Forecasting, and especially Classification. Produced several computer books and research in the computer field. He can be contacted by email: aswan.sunge@pelitabangsa.ac.id and orcid https://orcid.org/0000-0003-4296-9824.

Spits Warnars Harco Leslie Hendric, is a Professor in Computing at Bina Nusantara University, He is served as head of Information Systems concentration at Doctor of Computer Science, Bina Nusantara University. He did Bachelor's degree in Computer Science in the Information Systems field from STMIK Budi Luhur, Indonesia, and continued his Master's degree in Computer Science with a major in Information Technology at the University of Indonesia. His Ph.D. Computer Science was done at The Manchester Metropolitan University, Manchester, United Kingdom and his PhD was funded by the Directorate General of Higher Education, Ministry of Education and Culture, Republic of Indonesia (DIKTI) scholarship. His email: spits.hendric@binus.ac.id and orcid at https://orcid.org/0000-0002-5942-417X.

Vol. 10, No. 4, December 2024, pp. 747-762

Dendy K. Pramudito, Graduated Ph.D in Business Management from BINUS University in 2021. Currently is assigned as Director of Technology and Operation at Artajasa Pembayaran Elektronis, also actively as lecturer of information system and digital business in few universites. His research interests mainly in information system and digital business. Further discussion can be through email doktor.haji.dendy@pelitabangsa.ac.id and https://orcid.org/0000-0003-3079-8319.