RESEARCH PAPER

A Novel MCSUT Technique based on FastText Embedding for Improving Multi-URL Classification and Cybersecurity Performance

Zafar Ali[-1,]*, Siti Sophiayati Yuhaniz-[2,] , Wan Noor Hamiza-[3,], Jawaid Ahmed Siddiqui-[4,], Noureen-[5,] , Husham M. Ahmed[6]

[1]*Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur Malaysia*
[2]*Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur Malaysia*
[3]*Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, Kuala Lumpur, Malaysia*
[4]*Department of Computer Science, Sukkur IBA University, Sukkur 65200, Pakistan*
[5]*Department of Applied Computing and Artificial Intelligence, Universiti Teknologi Malaysia, Johor,*
[6]*College of Engineering, University of Technology, Bahrain, Kingdom of Bahrain,*
[1] ali.z@graduate.utm.my [2] sophia@utm.my; [3] wannoorhamiza@utm.my; [4] jawaid@iba-suk.edu.pk; [5] noureen@graduate.utm.my; [6]hmahmed@utb.edu.bh*

ABSTRACT
The exponential growth of web content requires efficient URL-based classification. Current methodologies utilize public URL classification datasets that fall into two categories, including DMOZ, Web Proxy Data, and WebKB, which are considered a general category. Other dataset categories, such as phishing, OpenPhishing, URLNet, Web Spam, and malicious, are part of the cybersecurity datasets. The datasets face challenges of class imbalance, noise, and ambiguity, which affect the performance of the URL classification models. To address these limitations, this study proposes an innovative multiple contextual semantic URL tokens (MCSUT) augmented technique that improves the quality of the URL classification dataset by reducing the noise and ambiguity contained in the URLs. The strength of the MCSUT technique mainly relies on its utilization of contextual and semantic URL tokens derived from neural word embedding techniques, such as WordNet, Word2Vec, and FastText, which are based on original tokens. This significantly enhances the ability of deep neural networks to comprehend and interpret these contextual and semantically rich tokens. This study presents a series of experimental results based on three-word embeddings using two datasets (DMOZ and phishing Datasets) and the development of data schemes for the DMOZ and phishing datasets, utilizing contextual and semantic tokens. The innovative multiple contextual semantic URL tokens (MCSUT) based on FastText neural word embeddings have outperformed previous studies, achieving a 0.8625 F1 score compared to WordNet, Word2Vec embeddings, and baselines, and achieved an F1 score of 0.99% on the phishing dataset...

Keywords: URL Classification; Cybersecurity; FastText Embedding; Deep Neural Networks; Data Augmentation.
How to cite this article: Ali Z, Yuhaniz SS, Hamiza WN, Siddiqui JA, Noureen, Ahmed HM., A Novel MCSUT Technique based on FastText Embedding for Improving Multi-URL Classification and Cybersecurity Performance .Int J Drug Deliv Technol. 2026;16(3s): 612-624; DOI: 10.25258/ijddt.16.3s.78
Source of support: Nil.
Conflict of interest: None

INTRODUCTION

The rapid growth of the internet has driven the need for effective URL classification systems to organize and filter online content for purposes like cybersecurity, search engine optimization, and personalized recommendations. According to Statista, the internet currently hosts over one billion registered websites, with thousands of new web pages being published daily. This unprecedented scale of digital content poses significant challenges in web management, optimization, and enforcing cybersecurity protocols. In response, the research community has proactively addressed these challenges by publishing open-access datasets and contributing to an ever-growing body of literature that advances URL classification methodologies. This collaborative effort seeks to improve these systems' accuracy, efficiency, and scalability. Some studies were conducted in the initial stages of URL classification research [55-60], primarily focusing on extracting features from web content [1-3]. This approach, however, proved to be computationally intensive and time-consuming. While comprehensive, it presented significant scalability challenges in the face of rapidly growing internet resources. Subsequent research paradigms shifted towards more efficient methodologies, prioritizing feature extraction directly from URLs [4, 5], [61-63].

The machine learning classifiers, including Logistic Regression, Random Forest, Support Vector Machines, K-Nearest Neighbors (KNN), and Multinomial Naive Bayes, were frequently employed to enhance the accuracy of the URL classification models through various feature engineering techniques [6-9]. Additionally, most studies in cybersecurity have predominantly concentrated on binary classification rather than addressing multiple URL classification, although this depends on the research objectives of each study [10, 11]. The multiple classification Khare, Bhandari and Murthy [12] leveraged web page meta-data, utilizing 10 gigabytes of proxy data for binary classification, achieved 81% accuracy with Naive

**Author for Correspondence: ali.z@graduate.utm.my*

Bayes using Term Frequency & Document Frequency. Abdallah and de La Iglesia [13] achieved 82.44% accuracy on the DMOZ dataset (200,000 entries) using NB and SVM with All-Grams, improving to 82.72% with N-Grams LM, and 66.5% on WebKB (4,167 entries) with NB and SVM using binary classification. Binary classification was also commonly reported in these studies [14-19].

Table 1 refers to word-based multiple-gram models; therefore, a single word is insufficient to determine the class of a URL, so the recommended two words that function as one word lead the classifier to a specific dataset category.

Table 1. N-Grams Shawon, Zuhori [20]

N-gram range (1,2)	Occurrence
computer	**1**
org	**1**
http computer	**1**
computer org	**1**

Another attempt [5, 13, 21] was made to improve the quality of features of the URL classification datasets by implementing character embedding and word embedding to capture all features at the most granular level, capturing patterns in character sequences that might signify malicious intent, specific content types, or domain categories. Moreover, it can establish a relationship between "irregular words" and regular words by comparing their alphabetical order.

URL classification, which encompasses both general and phishing categories, frequently contains ambiguous and misclassified data [22, 23]. Furthermore, the phishing category has more noise compared to the general URL category.

Table 2. Ambiguous and Misclassification Data Found In URLs

URL	Reason to Ambiguity	Category
‘http://lacasta.iespana.es’	Spanish website; content unclear without translation.	‘adult’
‘http://www.b4uhost.com’	Hosting site; unclear if used for adult content.	‘adult’
‘http://www.kidzplayzone.com’	Likely related to kids or games.	‘kids’
‘http://www.solarattic’	Could be confused for Home & Garden or even Science	‘business’
‘http://www.oilnergy.com’	Sounds scientific because of the word "energy"	‘business’
‘http://www.homedesigns.com’	It could be related to home improvement or interior design.	‘home’
‘http://www.gardenbliss.org’	Focuses on gardening, could be confused with home decor.	‘home’
‘http://www.dailybuzz.com’	Generic name suggests news or current events.	‘news’
‘http://www.globalupdates.net’ ‘https://bankofamerica.alerts-login.com/auth’ ‘http://drive.google.com.secure-folder-share.xyz/login’	Sounds like a news or updates site, but content unclear. Banking phishing-unclear content Google drive phishing-confusing URL	‘news’ ‘Phishing’ ‘Phishing’

In conclusion, the ambiguity of URLs, as highlighted in Table 2, underscores the inherent challenges in web content classification, particularly when addressing linguistic, contextual, and categorical uncertainties. URLs such as ‘http://www.solarattic’ and ‘http://www.oilnergy.com’ are actually related to the ‘business’ category, but sound like ‘home’ or ‘science’. Furthermore, some categories, such as ‘adult’, ‘phishing’, and others, have more noise compared to others. These factors pose more challenges to classification models. The prior studies have struggled to

achieve the performance specific to the challenging categories. Therefore, to address this critical gap, this study introduces the Multiple Contextual Semantic URL Tokens (MCSUT) technique, which generates the three URL data schemes based on WordNet, Word2Vec, and FastText. The MCSUT technique enhances the classifier's ability to comprehend URL tokens by establishing strong contextual and semantic bonds among them, enabling more accurate prediction of specific URL classes. The FastText data scheme is more effective than WordNet and Word2Vec because FastText relies on Character-level n-grams + word comparison to Word2Vec based on word-level [24, 25]. The URL data is more complex, with short text, non-meaningful words, and special characters being common [22, 26]. Consequently, the data scheme leveraging FastText word embedding using the MCUST technique demonstrates superior efficacy compared to alternative methods for URL classification across both general and cybersecurity datasets, including DMOZ, WebKB, and Phishing, as elaborated in the subsequent section. The principal contributions of this study are outlined as follows.

1) To develop a novel Multiple Contextual Semantic URL Tokens (MCSUT) technique based on FastText to address the challenges of noise, ambiguity, and misclassification prevalent in URL classification datasets.
2) To compare the effectiveness and performance of the WordNet, Word2Vec and FastText Embedding using URL Classification of datasets using the same deep neural network

3) The MCSUT technique based on FastText neural word embedding is more effective on the phishing dataset compared to general URL classification datasets

The rest of the paper is structured as follows: Section 2 presents the literature review. Section 3 demonstrates the class representation of the DMOZ dataset and its URL complexity structure. Section 4 discusses the details of the proposed MCSUT technique. Section 5 covers the results and discussion, followed by the conclusion and future work direction in Section 6.

2. Literature Review

URL classification has emerged as a key element in various applications, including web page categorization and malicious phishing content detection, drawing significant attention from the research community. In the early stages of URL classification research, studies emphasised feature extraction from web content [27] and [5]. However, subsequent research has predominantly shifted towards extracting features from URLs [4, 5, 21, 28, 29]. URL-based approaches were applied to both traditional machine learning and deep learning models [22, 28-34] and [54]. Kan et al [4, 35] utilized a complex segmentation or expansion technique to extract features from the URL. Baykan, Henzinger [36] recommended several methods using tokens and n-grams (n = 4 to 8) features. Rajalakshmi and Aravindan [28] suggested an experiential strategy for feature selection based on a dictionary.

Table 3. Literature Review URL Classification General and Cyber Security with limitations

References	Dataset	Classifier	Results and Limitations
Khare, Bhandari and Murthy [12],2014	Gigabytes of proxy data	NB	81% with Binary Class
Abdallah and de La Iglesia [13] ,2018	DMOZ, WebKB	SVM, LM	82.44% and 82.72% with Binary Class
Rajalakshmi, Tiwari [22]	DMOZ, WebKB	NB	Lower Accuracy 56% and 60%
Kexin, Liang [37]	DMOZ Dataset	CNN and RNN, URL Net	81.23% with soft 6 classes 50% with Hard Classes
Kurt and Demirel [26],2022	DMOZ Dataset	CNN, LSTM,GRU	0.7800,0.7600, 0.7700. Limited dataset
Weedon and Denholm-Price [38],2017	Phish Tank Dataset	RF	86.9% with Binary Class
Giri and Banerjee [39]	Phishing Dataset	Ensemble Model	F1-score: 97.07%, Bi-classification, can be improved
Pathak and Shrivas [40]	Phishing Dataset	RF Model with Feature Selection Using PSO	F1-score: 98.69%, Bi-classification, can be improved
Ghalechyan, Israyelyan [41]	Phishing Dataset	BERTY, Transfer learning techniques	Accuracy 97%, Bi-classification, can be improved
Kumar, Muthusamy and Jerald [42]	Phishing Dataset	CNN & RNN	F1-Score: 96.7%, Bi-classification, can be improved

According to Table 3, URL classification research has made progress but faces critical gap. The research with multiple URL classification have lower accuracies compared to binary URL classification. Over-reliance on binary classification, especially in phishing datasets, limits applicability to complex, multi-class scenarios. Variability in classifier performance—from Naive Bayes to advanced neural networks like CNN and BERT—reveals a lack of standardized methodologies. Many studies were struggled to deal with complex data reducing generalizability. While some models achieve high F1-scores, they often fail to address noise, ambiguity, and misclassification.

The prior studies offered several methods to augment the dataset. Data noising augmentation technique to regularize sequence models based on neural networks[43]. Easy data augmentation (EDA) consists of four basic operations: random insertion, synonym replacement, random swap, and random deletion [44]. By intelligently applying these four fundamental operations, one can significantly enhance the diversity and quality of training datasets. Moreover, in 2021, using a similar approach, it employed the random insertion of punctuation marks into the original text and showed better performance than the applied methods in 2019 [45]. The work by authors Connor, Shorten, Khoshgoftaar and Furht [46], explores a variety of data augmentation methodologies, including symbolic augmentation, rule-based augmentation, graph-structured augmentation, Mix-up augmentation, feature space augmentation, neural augmentation, back-translation augmentation, style augmentation, label augmentation, and generative data augmentation. The effectiveness of a particular data augmentation technique depends on the specificity of the data. Data augmentation is effective not only on image data but also on text and audio data.

The current proposed technique is based on data augmentation and word embedding methodologies. Word embeddings are foundational in natural language processing (NLP), representing words as real-valued vectors within a finite-dimensional space [38]. This approach ensures that semantically similar words are mapped to analogous vector representations, capturing their contextual and relational nuances. Word2Vec, GloVe, and FastText are the most widely used embedding techniques, each offering unique strengths tailored to specific tasks [47]. The Word2Vec excels in capturing semantic relationships through its continuous bag-of-words (CBOW) and skip-gram architectures, while the GloVe leverages global word co-occurrence statistics to generate embedding's [48].

FastText, developed by Facebook AI Research (FAIR), has emerged as a widely adopted tool for generating word embeddings, as demonstrated by the author [49] and [38]. Unlike traditional methods such as Word2Vec, FastText employs character n-grams to represent words, enabling it to capture morphological information effectively [53]. This unique approach allows FastText to handle out-of-vocabulary (OOV) words more precisely, making it particularly valuable in scenarios where vocabulary

coverage is incomplete or dynamic [50, 51]. The most important FastText feature, based on character n-grams, is practical during the word embedding training of the URLs because of the short text, non-meaningful words, special characters, symbols, and numbers commonly found in URLs [22, 52].

The researchers often struggled to achieve satisfactory performance in multi-class URL classification into the challenging categories 'adult', 'science' ,'kids' and 'phishing' due to noisy data (shortened string lengths, non-meaningful words), ambiguity, and misclassification. To tackle these challenges, this study presents data schemes based on MCSUT technique, using the WordNet, the Word2Vec, and the FastText word embedding approaches.

3. Dataset for URL Classification

This study focuses on the URL Classification dataset, the DMOZ dataset in the general category, and the phishing dataset in the cybersecurity category, all of which are publicly available datasets. The DMOZ dataset comprises 1.5 million URLs distributed across fifteen distinct categories. A detailed breakdown of these categories is provided in Table 4. This table illustrates that the dataset is imbalanced and contains several uneven samples in the training sets across multiple categories; it indicates that the categories 'arts,' 'society,' and 'business' hold a larger volume of data compared to other classes. The cybersecurity dataset is categorized as phishing, as shown in Table 5; the 'legitimate' category has a higher data representation than the 'phishing' category. The additional phishing dataset is considered a noisier and complex dataset compared to the general URL datasets. The URL classification dataset is also significantly more complex than a natural English word corpus due to shortened string lengths, noisy data, and ambiguity. Structural variability adds to the complexity, as URLs can vary widely in their composition, including paths, query parameters, and fragments.

Table 4. DMOZ Dataset Class Wise Representation

Serial No	Category Name	Number URLs	Overall Representation
1	ARTS	252401	16.23
2	SOCIETY	242488	15.59
3	BUSINESS	239555	15.40
4	COMPUTERS	117014	7.52
5	SCIENCE	109758	7.06
6	RECREATION	106099	6.82
7	SPORTS	100971	6.49
8	SHOPPING	94980	6.11
9	HEALTH	59887	3.85
10	REFERENCE	57636	3.71
11	GAMES	56287	3.62
12	KIDS	46002	2.96
13	ADULT	35100	2.26
14	HOME	28165	1.81
15	NEWS	8947	0.58

The phishing datasets also have challenges; the legitimate or benign category has a more representation of the data compared to phishing or malicious data. Adding this legitimate or benign category has more noisy data regarding the normal categories as mentioned in Table 4.

Table 5. Phishing Dataset Class Wise Representation

Serial No	Category Name	Number URLs	Overall Representation
1	Legitimate	275932	77.33
2	Phishing	80862	22.66
Total Number of Samples - 356794			

To address these challenges, this study proposes the Multiple Contextual Semantic URL Tokens (MCSUT) technique, designed to develop data schemes that significantly reduce noise, ambiguity, and misclassification compared to the original datasets. By leveraging advanced contextual and semantic analysis, the MCSUT technique enhances the quality and reliability of URL classification, offering a more robust solution for both general and cybersecurity applications.

4. Proposed Technique

This section outlines the proposed technique developed in the study. The data augmentation strategy and the classification model is described in detail. The proposed technique, referred to as Multiple Contextual Semantic URL Tokens (MCSUT), is based on a data augmentation approach that effectively utilizes contextual and semantic URL tokens.

These enriched tokens enhance the deep neural networks' capacity to better understand and interpret URL components, thereby improving overall model performance. The MCSUT technique leverages three advanced neural word embeddings: WordNet, Word2Vec, and FastText. Utilizing this approach, distinct data schemes—Simple, WordNet, Word2Vec, and FastText—have been developed. Significantly, the WordNet data scheme, the Word2Vec data scheme, and FastText data scheme incorporate contextual and semantic URL tokens that improve the data quality of the URL.

Figure 1 illustrates the workflow for developing data schemes using the Multiple Contextual Semantic URL Tokens (MCSUT) technique. The effectiveness of these dataset schemes is subsequently evaluated and validated using a consistent deep neural network, specifically the Bidirectional Long Short-Term Memory (BiLSTM) model. BiLSTM is particularly effective for contextual and semantic data because it can process sequential information in both forward and backward directions, more comprehensively capturing dependencies and relationships within the data. This bidirectional approach enables the model to comprehend the context and semantics of URL

tokens more effectively, thereby enhancing its ability to interpret complex patterns and nuances. The original dataset serves as the foundational input for the Multiple Contextual Semantic URL Tokens (MCSUT) technique, which constructs data schemes utilizing three distinct word embeddings: WordNet, Word2Vec, and FastText. The output of the MCSUT technique includes the Simple data scheme, the WordNet data scheme, the Word2Vec data scheme, and the FastText data scheme. These schemes are subsequently validated using a BiLSTM layer to assess their effectiveness and performance. Employing the same BiLSTM layer for validation ensures a standardized evaluation process, allowing for a fair comparison of the data schemes under identical experimental configurations.

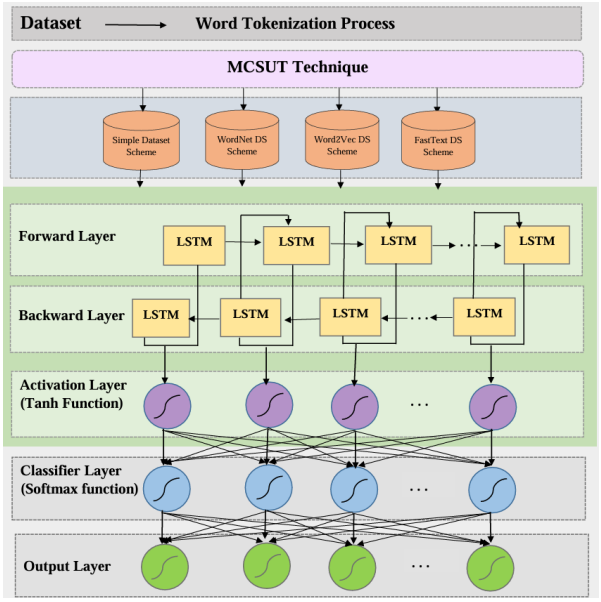

Figure 1. Workflow for Developing and Validating URL Data Schemes

4.1 Bidirectional Long Short-Term Memory Network

Unlike traditional machine learning's single-layer structure in deep learning architectures, operations move through multiple layers with different levels of complexity and many parameters. Each layer processes and transforms outputs using non-linear activation functions to identify critical data features. This study introduces Bi-LSTM-based multi-class classification deep learning models. The fundamental concepts and mathematical formulations of the Bi-LSTM architecture are detailed below. Recurrent Neural Networks (RNNs) have demonstrated strong performance in modeling short-term sequential dependencies. However, they encounter difficulties capturing long-term dependencies due to vanishing or exploding gradients.

To address these limitations, the Long Short-Term Memory (LSTM) architecture was developed as a specialized variant of RNNs. LSTM networks can maintain long-term dependencies by incorporating a memory cell structure regulated by three gates: input, output, and forget gates. While the overall structure of LSTM resembles standard RNNs, its neuron architecture is designed to effectively preserve information over more extended sequences. The Bidirectional LSTM (Bi-LSTM) architecture enhances the standard LSTM by connecting two LSTM layers in opposite temporal directions. This configuration enables the model to capture contextual information from both past and future states in a sequence, thereby producing a more comprehensive representation of the input data. (See Fig.2). The mathematical representation of a single LSTM memory cell is as follows:

$$f^t = \sigma(W_{xf}x^t + W_{hf}\,h^{t-1} + b_f) \quad (1)$$

$$i^t = \sigma(W_{xi}xt + W_{hi}h^t + bi) \quad (2)$$

$$c^{\sim} = tanh\,(W_{xc}\,x_t + W_{hc}h^{t-1} + b_c) \quad (3)$$

Update state

$$c_t = f_t \otimes c_t - 1 + i_t \otimes c^{\sim}\,t \quad (4)$$

$$o_t = \sigma(W_{xo}X_T + W_{hc}h^{t-1} + b_o) \quad (5)$$

$$h_t = o_t \otimes tanh\,(c_{t)} \quad (6)$$

i_t, f_t, c_t, o_t, and h_t represent the, input gate, forget gate, cell state, output gate, and concealed state, respectively. W represents the weight matrix linked to each layer, while b denotes the bias. The initial step in the LSTM process is to identify which information is unnecessary and should be removed from the cell state, a decision made by the sigmoid layer, also known as the forget gate. Following this, the LSTM determines what new information should be added by utilizing two layers: the input gate, a sigmoid layer that decides which values to update, and a tanh layer generates a vector for new candidate values to be stored in the cell state C~t. As equation (4) outlined, the previous cell state c_{t-1} is updated to the new cell state c_t. A sigmoid laye[0, 1] is then applied to determine which portion of the cell state will contribute to the outputo_t, this output is element-wise multiplied by a tanh layer that processes the cell state, with the tanh function limiting values to the range of -1 to 1, thereby producing the final hidden output.

Generation of URL Contextual and Semantic Tokens

Generating contextual tokens primarily aims to expand the URL token set while preserving the URL's integrity, as more relevant and closely related tokens enhance the classifier's performance. Studies have explored the generation of synonyms using WordNet, Word2Vec, and FastText methods, which have proven effective in enriching datasets and improving the performance of classification models. Word2Vec, a word embedding technique, captures contextual information by mapping words into a high-dimensional space, enabling the identification of semantically related terms—for instance, associating "shop" with "store" or "purchase." WordNet, a comprehensive lexical database, groups words into synonym sets (synsets) representing distinct concepts [37], making it particularly useful for analysing paths and query parameters within URLs. FastText, developed by Facebook AI, represents words as combinations of character-level sub-word units (n-grams), offering robust handling of morphological variations. Fig. 1 and Fig. 2 illustrate how the MCSUT technique, leveraging Word2Vec, WordNet, and FastText, is used to generate a contextual and semantically rich dataset scheme, such as the WordNet data scheme, the Word2Vec data scheme, and the FastText data

scheme, using the original DMOZ dataset and phishing dataset, which are publicly available.

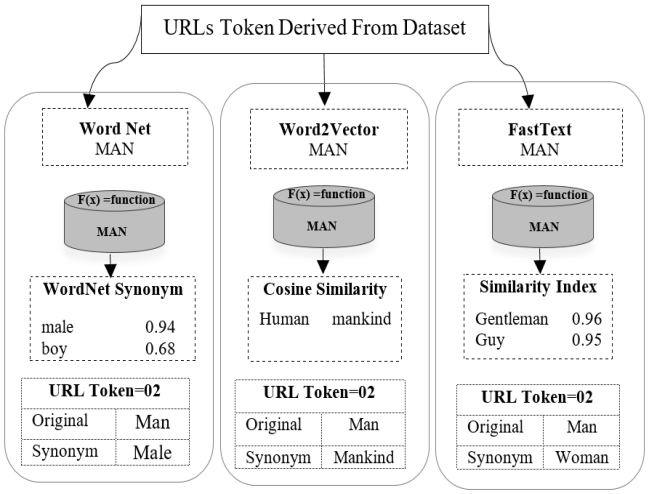

Figure 2. Process of Contextual URL Tokens

WordNet Lexical URL Tokens

It is a comprehensive and widely used lexical database for the English language, which makes it an excellent resource for finding synonyms. By incorporating synonyms, you can improve the classification accuracy. URLs might use different words with similar meanings, and WordNet can help identify these variations. WordNet can aid in better understanding the English words used in URLs, especially when dealing with domain-specific terminology Table 6.

Table 6. WordNet URL Tokens Generated by MCSUT Technique

MAN	STORY	PROFESSION	SEXUAL	PHISHING-security
human	tale	profession	sexual	signin
mankind	report	professing	intimate	authenticate
adult male	history	business	none	validation
human being	fib	none	none	password
humankind	news	none	none	credentials
FRIEND	**POWERFUL**	**PHISHING-Threat**	**CLOTHES**	**TREASURES**
acquaintance	mightily	warnning	clothes	cherish
supporter	powerful	suspended	garment	value
ally	muscular	blocked	encloth e	appreciate
booster	potent	deactive	dress	treasure
friend	right	pending	fit out	hoarded wealth

Table 6 shows the URL tokens of WordNet; all URL tokens will not be part of the data scheme; the first two tokens will be part of the data scheme; however, the Word2Vec and FastText URL tokens will be included in the data scheme based on similarity index values.

4.4 Word2Vec Contextual and Semantical URL Tokens

Word2Vec offers distinct advantages over WordNet for URL classification datasets, including the DMOZ and phishing data. Unlike WordNet, which relies on predefined lexical relationships, Word2Vec captures contextual word meanings by analyzing word co-occurrences in large text corpora.

Table 7. Word2Vec URL Tokens Generated by MCSUT Technique

MAN	STORY	PROFESSION	SEXUAL	PHISHING-security
woman, 0.76	stories, 0.73	professions, 0.75	sex, 0.76	login, 0.84
boy, 0.72	tale, 0.68	noble profession, .70	sex intercourse, .72	verify, 0.79
teenager, 0.65	fable, 0.56	vocation, 0.58	sexual, 0.70	password, 0.81
teenage girl, 0.61	tales, 0.54	careers, 0.53	sexually, 0.68	Session, 0.88
girl, 0.51	press, 0.50		intercourse, 0.67	security, 0.89
FRIENDS	**POWERFUL**	**PHISHING-Threat**	**CLOTHES**	**TREASURES**
acquaintances,.73	powerful, 0.71	Alert, 0.90	Clothing, 0.80	treasure, 0.78
friend 0.70	potent, 0.64	Immediate, 0.88	Underwear, 0.68	Archeological, 0.68
buddies, 0.70	influential, 0.54	urgent, 0.89	Underclothes, 0.66	relics, 0.68
pals, 0.68	powerful, 0.52	expired, 0.87	Garments, 0.64	artifacts, 0.67

relatives, 0.65	formidable, 0.51	failed,0.92	Jeans, 0.64	priceless, 0.67

Table 7 presents a sample of the URL token, which will be part of the data scheme based on the similarity index value if it exceeds 70. Moreover, dense vector representations that reflect semantic similarities, make it highly effective for interpreting complex URL tokens. Word2Vec excels in handling dynamic and diverse URL structures, as it learns from vast datasets, providing robust embeddings for both common and domain-specific terms. Its ability to model contextual relationships makes it a powerful tool for improving URL classification accuracy.

Fast Text Contextual and Semantic URL Tokens

In the context of URL classification, FastText word embedding demonstrates significant advantages over WordNet and Word2Vec, particularly when dealing with the complexities of URL datasets

.

Table 8. FastText URL Tokens Generated by MCSUT Technique

MAN	**STORY**	**PROFESSION**	**SEXUAL**	**PHISHING-security**
Guy, 0.88	Tale 0.90	Expert,0.86	Sensual,0.84	token,0.85
Gentleman, 0.82	Narrative 0.87	Specialist,0.80	Intimate,0.78	reset,0.92
Male, 0.78	Account 0.86	Skilled,0.78	Romantic, 0.76	update,0.90
Person 0.71	Plot 0.85	Qualified,0.76	Reproductive,0.72	privacy, 0.93
Dude 0.68	Chronicle 0.66	Experienced,0.72	Gendered, 0.68	protection,0.91
FRIEND	**POWERFUL**	**PHISHING-threat**	**CLOTHES**	**TREASURES**
Budy,0.70	Mighty, 0.90	risk,0.89	Clothing, 0.84	Riches,0.80
Pale,0.68	Strong, 0.88	attack,0.91	Garment,0.82	Valuables 0.76
Mate,0.66	Influential0.82	hazard,0.95	Outfit,0.78	Gems, 0.56
Booster,0.64	Potent, 0.84	malware, 0.93	Wardrobe, 0.74	Jewels 0.42

While WordNet relies on predefined lexical relationships and Word2Vec focuses on whole-word embeddings, FastText incorporates character-level sub-word information, uniquely suited for handling URLs' dynamic and intricate nature. Unlike Word2Vec, which processes words as single units, FastText breaks words into n-grams, effectively capturing morphological variations, abbreviations, and misspellings. FastText embedding is more effective for the phishing or malicious URL because of the character level attribute. FastText's sub-word approach ensures robust generalization across diverse and dynamic URL structures, enhancing the classifier's ability to understand and interpret contextual nuances. The Table 8 represented sample of the contextual URL token that is generated base on the original token, now it is the part of the data scheme based on similarity index value; if the URL token similarity index value is higher than 70 then it will be part of the data scheme and URL token must be identical.

The MCSUT technique, employing Word2Vec, WordNet, and FastText, constructs a contextually and semantically enriched data, including the WordNet data scheme, the Word2Vec data scheme, and the FastText data scheme, leveraging the publicly available original DMOZ dataset and phishing dataset. The efficacy of MCSUT technique lies in its ability to establish precise contextual and semantic URL token relationships, which critically influence the performance of classifiers, as their accuracy inherently depends on dataset quality.

5. Results and Discussion

This section presents a detailed analysis of the experiments conducted to evaluate the effectiveness of data schemes developed using the Multiple Contextual Semantic URL Tokens (MCSUT) technique. The schemes are constructed from URL tokens extracted from the original datasets of DMOZ and phishing, with three out of the four schemes incorporating contextual and semantic relationships among these tokens. To ensure transparency and consistency in the evaluation process, a series of experimental configurations were implemented using the same deep neural network, specifically the BiLSTM model.

Table 9. F1 Score Evaluation of MSCUT Data Schemes (DMOZ-Based) Using BiLSTM Classifier

Classes	**Simple**	**WordNet**	**Word2Vec**	**Fast Text Data Scheme**
Arts	0.816	0.808	0.8446	0.8736
Busine	0.843	0.791	0.8711	0.8651
Societ	0.822	0.811	0.8713	0.8990
Sports	0.978	0.954	0.9721	0.9767
Scienc	0.693	0.631	0.6438	0.6728
Comp	0.857	0.875	0.8873	0.9010
Shopp	0.954	0.902	0.9412	0.9451
Recre	0.925	0.893	0.9350	0.9352
Refere	0.812	0.879	0.9245	0.9186
Health	0.955	0.903	0.9472	0.9510
Games	0.717	0.784	0.7458	0.8059
Home	0.930	0.724	0.923	0.9613
Adult	0.558	0.646	0.7431	0.7690
News	0.652	0.859	0.8637	0.9044
Kids	0.495	0.674	0.5896	0.6142
Avera	**0.815**	**0.827**	**0.8491**	**0.8625**

Table 10. Evaluation of MSCUT Data Schemes (Derived From Phishing Dataset) Using the BiLSTM Classifier

Classes	Simple Token	Word Net	Word2 Vec	FastTe xt Data
Phishing	0.9001	0.9654	0.9758	0.9949
Legitimat	0.9156	0.9449	0.9696	0.9842

The experimental results demonstrate the effectiveness of the BiLSTM classifier using data schemes based on DMOZ and the phishing dataset for URL classification tasks. The FastText data scheme consistently outperformed WordNet and Word2Vec, achieving the highest average accuracy of 0.8625 based on the DMOZ dataset, as shown in Table 9. However, the FastText data scheme based on the phishing dataset demonstrated superior performance, achieving an F1 score of 0.9949 in the phishing category and an F1 score of 0.9842 in the legitimate category, as illustrated in Table 10.

The MCSUT technique is based on FastText's ability to capture sub-word information, which is particularly beneficial for handling noise, ambiguity, and morphologically complex URLs. By leveraging character n-gram embeddings, FastText enhances the semantic representation of tokens, enabling the BiLSTM classifier to better understand and interpret the contextual nuances of URLs. In contrast, WordNet's dataset scheme relies on predefined lexical relationships, while Word2Vec's dataset scheme focuses on whole-word embeddings, which limits their adaptability to the diverse and dynamic nature of URL datasets. The combination of FastText's data scheme with a deep neural network provides a robust and scalable solution for URL classification, including complex categories such as 'adult', 'kids', and 'phishing'.

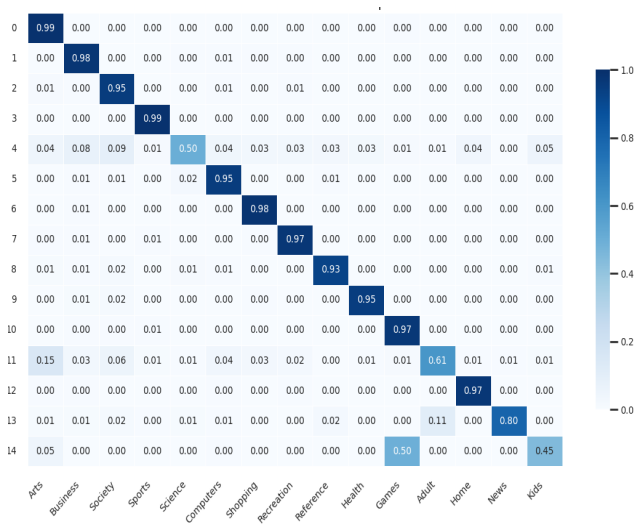

Figure 3. Confusion Matrix DMOZ Data Scheme using BiLSTM Model

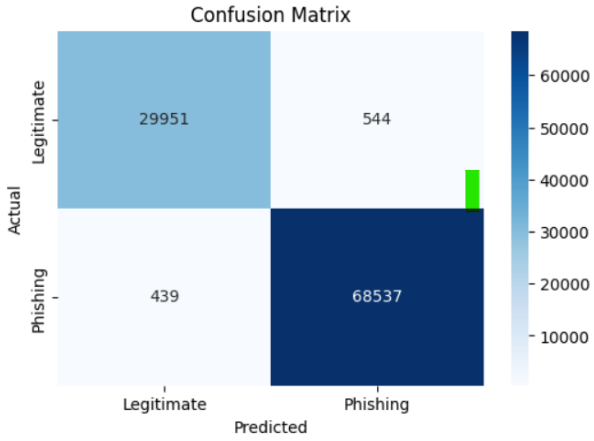

Figure 4. Confusion Matrix Phishing Data Scheme using BiLSTM Model

Referencing Fig 3, the Arts (0.9888), Business (0.9757), Sports (0.9892), and Shopping (0.9807) categories demonstrate the model's ability to classify URLs with minimal error. It also performs well in nuanced categories like Society (0.9524) and Health (0.9485), showcasing its capacity to handle complex, contextually rich URLs. However, the model faces challenges in ambiguous categories such as Science (0.5001) and Adult (0.6058), where overlapping domains and sensitive content reduce confidence. The Kids category (0.4463) also shows lower accuracy, likely due to the diverse nature of children's content. Additionally, Fig 4 demonstrates the model's higher performance when using the phishing data scheme; only a few phishing and legitimate URLs are misclassified. Overall, the BiLSTM classifier proves robust in capturing contextual and semantic nuances, making it a reliable solution for URL classification, though enhancements in ambiguous categories could further improve its real-world applicability.

Table 11. Comparison of the Four-Dataset Scheme with Baseline Studies

Research Studies	Approach	Training, Testing	Classes	F1Score
Abdallah and de La Iglesia [13]	SVM	Limited Dataset	15	82.72 %
Rajalakshmi, Tiwari [22],2020	RNN with GRU	46280 For Kids & Rest of URL. 80%, and 20%	Kids with rest of all	0.7792
	RNN with Bidirectional GRU			0.7962
	Bidirectional GRU with CNN			0.8204
	TextCNN	1.5 Million	15 Classes	0.500

Kexin, Liang [37],2021	TextCNN		6 Soft Classes	0.8107 *
	TextCNN		6 Hard Classes	0.7346
Giri and Banerjee [39]	Phishing Dataset	Ensemble Model	Binary Class	97.07 %,
Pathak and Shrivas [40]	Phishing Dataset	RF Model with Feature Selection Using PSO	Binary Class	98.69 %,
Kumar, Muthusamy and Jerald [42]	Phishing Dataset	CNN & RNN	Binary Class	96.7%
Kurt and Demirel [26],2022	CNN	120K, 30K	15 Classes	0.7800
	LSTM			0.7600
	GRU			0.7700
Current Study	Simple Tokens	1.5+ Million, 90K	DMOZ Dataset 15 Classes	0.8159
	MCSUT WordNet Data Scheme			0.8278
	MCSUT Word2Vec Data Scheme			0.8491
	MCSUT FastText Data Scheme			0.8625
	Simple Tokens	356794 Phishing Dataset	Phishing Category	0.9001
	MCSUT WordNet Data Scheme			0.9654
	MCSUT Word2Vec Data Scheme			0.9758
	MCSUT FastText Data Scheme			0.9949

According to Table 11, the current study demonstrates significant advancements in URL classification, outperforming several previous approaches. Using the DMOZ dataset, the study achieves an F1 score of 0.8625 with the FastText data scheme, surpassing traditional methods like SVM (82.72%) and CNN (0.7800). The Phishing dataset results are even more impressive, with the FastText data scheme achieving an F1 score of 0.9949, outperforming ensemble models (97.07%) and RF with PSO (98.69%). The study's use of BiLSTM with WordNet and Word2Vec also shows consistent improvement, with F1 scores of 0.9654 and 0.9758, respectively. These results highlight the effectiveness of leveraging advanced embedding's like FastText and BiLSTM for capturing contextual and semantic nuances in URLs. While previous studies like Kexin and Liang (2021) struggled with hard classes (0.7346), the current study's robust data schemes and deep learning models significantly enhance classification accuracy, setting a new benchmark for URL classification in both general and cybersecurity context.

6. Conclusion

The study demonstrates the effectiveness of the Multiple Contextual Semantic URL Tokens (MCSUT) technique combined with the BiLSTM classifier for URL classification. The FastText data scheme outperformed others, achieving an F1 score of 0.8625 on the DMOZ dataset and 0.9949 on the phishing dataset, thanks to its ability to capture sub-word information and handle noise and ambiguity. While excelling in categories like Sports (0.9767) and Shopping (0.9451), the model struggles with categories such as Science (0.6728) and Kids (0.6142). Compared to baseline studies, the MCSUT technique sets a new benchmark, surpassing traditional methods like SVM (82.72%) and CNN (0.7800). This approach offers a robust solution for URL classification, though further improvements in ambiguous categories could enhance real-world applicability.

Acknowledgment
The authors would like to gratefully acknowledge the Faculty of Artificial Intelligence, Universiti Teknologi Malaysia, and Kuala Lumpur, Malaysia.

Data Availability
The DMOZ dataset, a publicly available resource for URL classification, can be accessed on Kaggle at the following link:https://www.kaggle.com/datasets/shawon10/url-classification-dataset-dmoz. This phishing dataset is accessible at https://data.mendeley.com.

Authors' declaration
Conflicts of Interest: None. We hereby confirm that all the Figures and Tables in the manuscript are ours.
Authors' contributions
Zafar Ali contributed to the literature review, development, and implementation of the model and its features. Siti Sophiayati Yuhaniz and Wan Noor Hamiza worked on the research design and critically analyzed the results. Jawaid Ahmed Siddiqui assisted in reviewing the manuscript, while Noureen created the figures for the methodology section, and Husham M. Ahmed has worked on reviewing, editing and addressing the comments.

REFERENCE

1. Chakrabarti, S., M. van den Berg, and B. Dom, Focused crawling: a new approach to topic-specific Web

resource discovery. Computer Networks, 1999. 31(11): p. 1623-1640.
2. Pierre, J.M. Practical issues for automated categorization of web sites. in Electronic Proc. ECDL 2000 Workshop on the Semantic Web. 2000. Citeseer.
3. Chaker, J. and O. Habib, Genre Categorization of Web Pages, in Proceedings of the Seventh IEEE International Conference on Data Mining Workshops. 2007, IEEE Computer Society. p. 455–464.
4. Kan, M.-Y. and H.O.N. Thi. Fast webpage classification using URL features. in Proceedings of the 14th ACM international conference on Information and knowledge management. 2005.
5. Baykan, E., et al. Purely url-based topic classification. in Proceedings of the 18th international conference on World wide web. 2009.
6. 6Kaddoura, S., et al., A systematic literature review on spam content detection and classification. PeerJ Computer Science, 2022. 8: p. e830.
7. Murty, C. and P.H. Rughani, Dark web text classification by learning through SVM optimization. J Adv Inf Technol, 2022. 13(6).
8. Sánchez-Paniagua, M., et al., Phishing URL detection: A real-case scenario through login URLs. IEEe Access, 2022. 10: p. 42949-42960.
9. Yan, H., et al. Detecting malicious URLS using a deep learning approach based on stacked denoising autoencoder. in Trusted Computing and Information Security: 12th Chinese Conference, CTCIS 2018, Wuhan, China, October 18, 2018, Revised Selected Papers 12. 2019. Springer.
10. Ahmed, N., et al., Machine learning techniques for spam detection in email and IoT platforms: analysis and research challenges. Security and Communication Networks, 2022. 2022(1): p. 1862888.
11. Alkhodhairy, G. and K. Saleem, Machine learning algorithm for detecting suspicious email messages using Natural Language Processing NLP. Alexandria Engineering Journal, 2025. 128: p. 153-165.
12. Khare, S., A. Bhandari, and H.A. Murthy. Url classification using non negative matrix factorization. in 2014 Twentieth National Conference on Communications (NCC). 2014. IEEE.
13. Abdallah, T.A. and B. de La Iglesia. URL-based web page classification: With n-gram language models. in International Joint Conference on Knowledge Discovery, Knowledge Engineering, and Knowledge Management. 2014. Springer.
14. Darling, M., et al. A lexical approach for classifying malicious URLs. in 2015 international conference on high performance computing & simulation (HPCS). 2015. IEEE.
15. Mamun, M.S.I., et al. Detecting malicious urls using lexical analysis. in International Conference on Network and System Security. 2016. Springer.
16. Verma, R. and A. Das. What's in a url: Fast feature extraction and malicious url detection. in Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics. 2017.
17. Das, A., et al. Deep approaches on malicious URL classification. in 2020 11th International Conference on Computing, Communication and Networking Technologies (ICCCNT). 2020. IEEE.
18. Johnson, C., et al., Towards Detecting and Classifying Malicious URLs Using Deep Learning. J. Wirel. Mob. Networks Ubiquitous Comput. Dependable Appl., 2020. 11(4): p. 31-48.
19. Chang, W., F. Du, and Y. Wang. Research on malicious URL detection technology based on BERT model. in 2021 IEEE 9th International Conference on Information, Communication and Networks (ICICN). 2021. IEEE.
20. Shawon, A., et al. Website classification using word based multiple n-gram models and random search oriented feature parameters. in 2018 21st International Conference of Computer and Information Technology (ICCIT). 2018. IEEE.
21. Rajalakshmi, R. and C. Aravindan. Web page classification using n-gram based URL features. in 2013 fifth international conference on advanced computing (ICoAC). 2013. IEEE.
22. Rajalakshmi, R., et al., Design of Kids-specific URL Classifier using Recurrent Convolutional Neural Network. Procedia Computer Science, 2020. 167: p. 2124-2131.
23. Zieni, R., L. Massari, and M.C. Calzarossa, Phishing or not phishing? A survey on the detection of phishing websites. IEEE Access, 2023. 11: p. 18499-18519.
24. Lumbantoruan, R., et al. Analysis comparison of FastText and Word2vec for detecting offensive language. in 2022 IEEE International Conference of Computer Science and Information Technology (ICOSNIKOM). 2022. IEEE.
25. Luong, H.H., L.T.T. Le, and H.T. Nguyen. An approach for web content classification with FastText. in International Conference on Computational Data and Social Networks. 2023. Springer.
26. Kurt, M.S. and E.Y. Demirel, Web page classification with deep learning methods. Uludağ Üniversitesi Mühendislik Fakültesi Dergisi, 2022. 27(1): p. 191-204.
27. Shen, D., et al. Web-page classification through summarization. in Proceedings of the 27th annual international ACM SIGIR conference on Research and development in information retrieval. 2004.

28. Rajalakshmi, R. and C. Aravindan. Naive bayes approach for website classification. in International Conference on Advances in Information Technology and Mobile Communication. 2011. Springer.
29. Rajalakshmi, R. and C. Aravindan, A Naive Bayes approach for URL classification with supervised feature selection and rejection framework. Computational Intelligence, 2018. 34(1): p. 363-396.
30. Gupta, A. and R. Bhatia, Ensemble approach for web page classification. Multimedia Tools and Applications, 2021. 80(16): p. 25219-25240.
31. Gupta, A. and R. Bhatia, Knowledge based deep inception model for web page classification. Journal of Web Engineering, 2021. 20(7): p. 2131-2168.
32. Lee, J.-H., W.-C. Yeh, and M.-C. Chuang, Web page classification based on a simplified swarm optimization. Applied Mathematics and Computation, 2015. 270: p. 13-24.
33. Triplett, W.J., Addressing cybersecurity challenges in education. International Journal of STEM Education for Sustainability, 2023. 3(1): p. 47-67.
34. Yu, Y., Web page classification algorithm based on deep learning. Computational intelligence and neuroscience, 2022. 2022(1): p. 9534918.
35. Kan, M.-Y. Web page classification without the web page. in Proceedings of the 13th international World Wide Web conference on Alternate track papers & posters. 2004.
36. Baykan, E., et al., A comprehensive study of features and algorithms for URL-based topic classification. ACM Transactions on the Web (TWEB), 2011. 5(3): p. 1-29.
37. Kexin, X., et al., URL Classification with Deep Learning. 2021.
38. Weedon, M., Tsaptsinos, D., and J. Denholm-Price, Random Forest Explorations for URL Classification. In 2017 International Conference On Cyber Situational Awareness, Data Analytics And Assessment (Cyber SA), 2017(1-4).
39. Giri, S. and S. Banerjee, Ensemble Learning Approach for Phishing Website Detection Using an Optimal Greedy Stacking Model. Journal of The Institution of Engineers (India): Series B, 2024.
40. Pathak, P. and A.K. Shrivas, Development of Proposed Model Using Random Forest with Optimization Technique for Classification of Phishing Website. SN Computer Science, 2024. 5(8): p. 1-20.
41. Ghalechyan, H., et al., Phishing URL detection with neural networks: an empirical study. Scientific Reports, 2024. 14(1): p. 25134.
42. Kumar, S.S., P. Muthusamy, and M.P.A. Jerald, A Hybrid Framework for Improved Weighted Quantum Particle Swarm Optimization and Fast Mask Recurrent CNN to Enhance Phishing-URL Prediction Performance. International Journal of Computational Intelligence Systems, 2024. 17(1): p. 251.
43. Xie, Z., et al., Data noising as smoothing in neural network language models. arXiv preprint arXiv:1703.02573, 2017.
44. Wei, J. and K. Zou, Eda: Easy data augmentation techniques for boosting performance on text classification tasks. arXiv preprint arXiv:1901.11196, 2019.
45. Karimi, A., L. Rossi, and A. Prati, AEDA: an easier data augmentation technique for text classification. arXiv preprint arXiv:2108.13230, 2021.
46. Shorten, C., T.M. Khoshgoftaar, and B. Furht, Text data augmentation for deep learning. Journal of big Data, 2021. 8(1): p. 101.
47. Selva Birunda, S. and R. Kanniga Devi, A review on word embedding techniques for text classification. Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020, 2021: p. 267-281.
48. Akhtar, M.S., et al., Improving word embedding coverage in less-resourced languages through multi-linguality and cross-linguality: a case study with aspect-based sentiment analysis. ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP), 2018. 18(2): p. 1-22.
49. Joulin, A., et al., Fasttext. zip: Compressing text classification models. arXiv preprint arXiv:1612.03651, 2016.
50. Nasution, N.A., E.B. Nababan, and H. Mawengkang. Comparing LSTM Algorithm with Word Embedding: FastText and Word2Vec in Bahasa Batak-English Translation. in 2024 12th International Conference on Information and Communication Technology (ICoICT). 2024. IEEE.
51. Rana, A., et al. Semantic Similarity Analysis using FastText. in 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC). 2024. IEEE.
52. AL-Jumaili, A.S.A. and H.K. Tayyeh, Character-Level Embedding using FastText and LSTM for Biomedical Named Entity Recognition. Scalable Computing: Practice and Experience, 2024. 25(6): p. 5258–5264-5258–5264.
53. Huspi, S. H. H., & Ali, Z. (2024). Sentiment analysis on roman urdu students' feedback using enhanced word embedding technique. Baghdad Science Journal, 21(2 (SI)), 0725-0725.
54. Refianti, R. and N. Anggraeni, Sentiment Analysis Using Convolutional Neural Network Method to Classify Reviews on Zoom Cloud Meetings Application Based on Reviews on Google Playstore. Int. J. Eng. Sci. Inf. Technol,

2023. 3(3): p. 7-16.
55. Zuhanda, M.K., et al., A Hybrid GDHS and GBDT Approach for Handling Multiclass Imbalanced Data Classification. 2025. 5(3): p. 51-57.
56. Ahmed QW, Garg S, Rai A, Ramachandran M, Jhanjhi NZ, Masud M, Baz M. AI-Based Resource Allocation Techniques in Wireless Sensor Internet of Things Networks in Energy Efficiency with Data Optimization. Electronics. 2022; 11(13):2071. https://doi.org/10.3390/electronics11132071
57. N. Zaman, T. J. Low and T. Alghamdi, "Energy efficient routing protocol for wireless sensor network," 16th International Conference on Advanced Communication Technology, Pyeongchang, Korea (South), 2014, pp. 808-814, doi: 10.1109/ICACT.2014.6779072.
58. Jhanjhi, N.Z. (2025). Investigating the Influence of Loss Functions on the Performance and Interpretability of Machine Learning Models. In: Pal, S., Rocha, Á. (eds) Proceedings of 4th International Conference on Mathematical Modeling and Computational Science. ICMMCS 2025. Lecture Notes in Networks and Systems, vol 1399. Springer, Cham. https://doi.org/10.1007/978-3-031-91005-0_43
59. N. Jhanjhi, "Comparative Analysis of Frequent Pattern Mining Algorithms on Healthcare Data," 2024 IEEE 9th International Conference on Engineering Technologies and Applied Sciences (ICETAS), Bahrain, Bahrain, 2024, pp. 1-10, doi: 10.1109/ICETAS62372.2024.11119839.
60. Alkinani MH, Almazroi AA, Jhanjhi N, Khan NA. 5G and IoT Based Reporting and Accident Detection (RAD) System to Deliver First Aid Box Using Unmanned Aerial Vehicle. Sensors. 2021; 21(20):6905. https://doi.org/10.3390/s21206905
61. Z. A. Almusaylim, N. Zaman and L. T. Jung, "Proposing A Data Privacy Aware Protocol for Roadside Accident Video Reporting Service Using 5G In Vehicular Cloud Networks Environment," 2018 4th International Conference on Computer and Information Sciences (ICCOINS), Kuala Lumpur, Malaysia, 2018, pp. 1-5, doi: 10.1109/ICCOINS.2018.8510588.
62. S.H. Kok, A. Azween, NZ Jhanjhi, Evaluation metric for crypto-ransomware detection using machine learning, Journal of Information Security and Applications, Volume 55, 2020, 102646, https://doi.org/10.1016/j.jisa.2020.102646.
63. S. J. Hussain, U. Ahmed, H. Liaquat, S. Mir, N. Jhanjhi and M. Humayun, "IMIAD: Intelligent Malware Identification for Android Platform," 2019 International Conference on Computer and Information Sciences (ICCIS), Sakaka, Saudi Arabia, 2019, pp. 1-6, doi: 10.1109/ICCISci.2019.8716471