NATURENGS
MTU Journal of Engineering and Natural Sciences
https://dergipark.org.tr/tr/pub/naturengs
DOI: 10.46572/naturengs.1097970

Research Article

Using and Comparing Machine Learning Techniques for Automatic Detection of Spam URLs

Muhammed YILDIRIM[1*]

[1]Department of Computer Engineering, Faculty of Engineering and Natural Sciences, Malatya Turgut Ozal University, Malatya, Turkey.

(**Received:** 03.04.2022; **Accepted:** 18.05.2022)

ABSTRACT: With the developing technology, the issue of cyber security has become one of the most common and current issues in recent years. Spam URLs are one of the most common and dangerous issues for cybersecurity. Spam URLs are one of the most widely used attacks to defraud users. These attacks cause users to suffer monetary losses, steal private information, and install malicious software on their devices. It is very important to detect such threats promptly and take precautions against them. Detection of spam URLs is mainly done by using blacklists. However, these lists are insufficient to detect newly created URLs. Machine learning techniques have been developed to overcome this deficiency in recent years. In this study, URL classification was made using different machine learning techniques. In the study, 9 different classifiers were preferred for URL classification. The performances of the classifiers were compared in the URL classification process. In addition, similar studies in the literature have been comprehensively examined and these studies have been discussed. In addition, since the preparation of datasets in the natural language processing process greatly affects the training of models, these steps are discussed in detail.

Keywords: Cyber Security, Machine Learning, NLP, URL Detection, Classifiers.

1. INTRODUCTION

Developing technology has led to the emergence of different fields. Natural language processing is one of these current areas. Recently, it is a new field in that researchers have extensively researched and developed various applications. Naturel Language Processing (NLP) is a sub-branch of artificial intelligence that aims to understand, analyze, interpret and produce the natural language humans use with the developed systems. NLP brings together the steps of linguistics, artificial intelligence, computer technologies, statistics, and data processing [1, 2]. Evolving technology has led to the emergence of different fields. Natural language processing is one of these current areas. Recently, it is a new field in that researchers have extensively researched and developed various applications. NLP is a sub-branch of artificial intelligence that aims to understand, analyze, interpret and produce the natural language humans use with the developed systems. NLP brings together the steps of linguistics, artificial intelligence, computer technologies, statistics, and data processing [3].

Natural language processing is used in different fields. It is used in many areas, such as text summarization [4], sentiment analysis [5], correction of typos [6], translation systems [7], information extraction [8], and natural language production. NLP was also used in this study to

[*]Corresponding Author: muhammed.yildirim@ozal.edu.tr
ORCID number of authors: [1] 0000-0003-1866-4721

detect Spam URLs. Spam URLs appear as unwanted pop-ups and links that we encounter on the internet every day. Spam URLs can cause users to experience financial losses and steal their private information. In addition, spam URLs can change the ranking of searched pages and negatively affect network traffic. There are improved methods for detecting these URLs. The most commonly used method is the blacklist method, which keeps URL records. The biggest disadvantage of the blacklist method is that it cannot detect newly created spam URLs. Studies have been carried out in recent years to cope with this problem by using artificial intelligence and machine learning methods [9, 10]. In the machine learning approach, models are first trained using training data. These models, which are then introduced, classify newly emerged URLs as spam or normal. This approach is preferred to eliminate the problem of not detecting new sites in the black list method.

The number of words in natural languages is relatively high and words can have more than one meaning. It is a complicated process for machines to understand the different meanings of words. Therefore, machine understanding of natural languages is a difficult process. The increasing importance of NLP with each passing day lies in the acceleration of artificial intelligence studies with the developing technology.

NLP is an up-to-date field that enables the communication between humans and machines. NLP is a popular sub-branch of artificial intelligence that aims to understand natural languages by machines, analyze these languages and make inferences from them [11]. Machines' understanding of people's language will solve many real-life problems and allow people to find a more comfortable space.

In this study, for the machines' URL data to be processed, it must first be converted into a format that the machines can understand. In the study, firstly, text preprocessing steps were applied to the data in the dataset. This step has a great impact on the performance of the models. Then, feature maps were obtained using the bag of words method. The feature maps obtained in the last step were classified into different classifiers.

1.1.Related Works

Spam URLs are attacks that put both individual users and companies in a difficult position. There are various studies in the literature to minimize these attacks.
Do Xuan et al. tested machine learning methods on two different datasets to classify URL addresses in their study. Two different classifiers were used in this study. The researchers also ran the models in 10 and 100 iterations in this study and compared the results. Accuracy values of 93.39% and 90.70%, respectively, were obtained in each dataset in 100 iterations of the SVM classifier [12]. The researchers stated that they prioritized time and accuracy in this study.

Patgiri et al. used machine learning methods to detect malicious URLs in their study. In the study, URLs were classified as good and bad. Two different classifiers were used in the study and the dataset was divided into train and test at different rates. The results obtained by separating the dataset as train and test at different ratios were compared. The researchers stated that their accuracy value in the Random Forest classifier was higher than in the SVM classifier [13].

Jain et al. used URL addresses for phishing detection. Researchers stated that they have developed a new system to prevent phishing in their studies and that the system they developed

works with 14 features. It has been seen that SVM and Naive Bayes are used in the proposed system, and the SVM classifier is more successful. The success rate of this proposed system in detecting phishing has been 90% [14].

Joshi and colleagues explained that most of today's cyberattacks and scams originate from malicious websites. They stated that malicious URLs are delivered to users in different ways and that these URLs cause different harm to users. This study observed that machine learning methods were used to detect malicious URLs and an average of 92% accuracy value was obtained from 5 different data used for testing [15].

In this study, Goh et al. used 2 different datasets to detect spam URLs. In this study, the researchers obtained accuracy values by using different classifiers. In this study, the most successful results were obtained in the RF classifier and these accuracy values were 93.7% and 85.2% in each dataset, respectively [16].

Sun et al. used three different datasets for URL classification in their study. In this study, they used different machine learning techniques for spam detection. These nine machine learning techniques they use are frequently used in the literature. They obtained the highest F-measure value in the RF classifier in the first dataset and the C5.0 classifier in the second and third datasets. These values are 82.19%, 87.48% and 91.90%, respectively [17].

1.2.Contributions and Innovation

NLP has become one of the most popular topics in information technology in recent years. Because the developing technology has brought large amounts of digital data with it, it is of great importance to process these data and draw meaningful conclusions from them. It is difficult for machines to process data, especially in natural languages. In this study, URLs that put users and companies in a difficult situation regarding cyber security have been identified. In this study, for the classification of URLs, the data was first prepared in a way that the classifiers could understand. At this stage, data cleaning and editing steps are available. This is a step that closely concerns the performance of the models. After this step, the bag of words matrix was obtained. 9 different machine learning methods were used to classify the URLs in the dataset. Performance metrics obtained in 9 different classifiers are discussed in detail.

1.3.Flow of Paper

Organization of the paper; In the first part, general information and related studies are given, and in the second part, the background part is provided. This section examines the dataset, data cleaning methods, and classifiers used in the study. In the third section, the experimental results and the last section, the results are discussed.

2. BACKGROUND

In this section, the dataset used in the study was examined, the data cleaning and data preparation stages were detailed and the techniques used in the study were discussed. A summary representation of the proposed model is given in Figure 1.

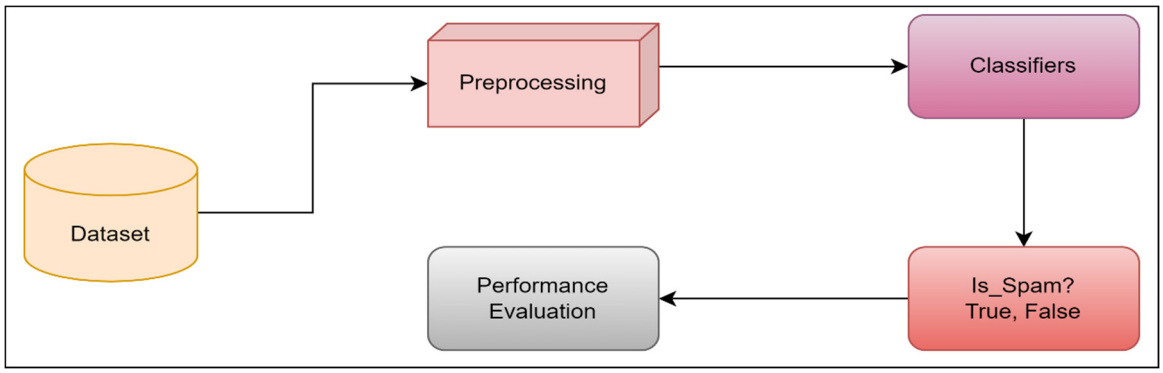

Figure 1. Summary of the proposed model

2.1. Dataset

The dataset used in the study was taken from Kaggle. This dataset consists of 148303 data in total. 101021 of these data are not malicious spam URLs [18]. The remaining 47282 are spam URLs. Sample data from the dataset are given in Figure 2.

index	url	is_spam
0	https://briefingday.us8.list-manage.com/unsubscribe	true
1	https://www.hvper.com/	true
2	https://briefingday.com/m/v4n3i4f3	true
3	https://briefingday.com/n/20200618/m#commentform	false
4	https://briefingday.com/fan	true

Figure 2. Example URLs from the dataset

Machine learning methods cannot do URL classification directly on the text. This dataset, which consists of URLs, must first be prepared in a format that machine learning methods can understand. In the study, the steps in Figure 3 were performed before the data were classified, and the data were prepared in a format that the models could understand. This data cleaning and data preparation process greatly impacts the performance of the models.

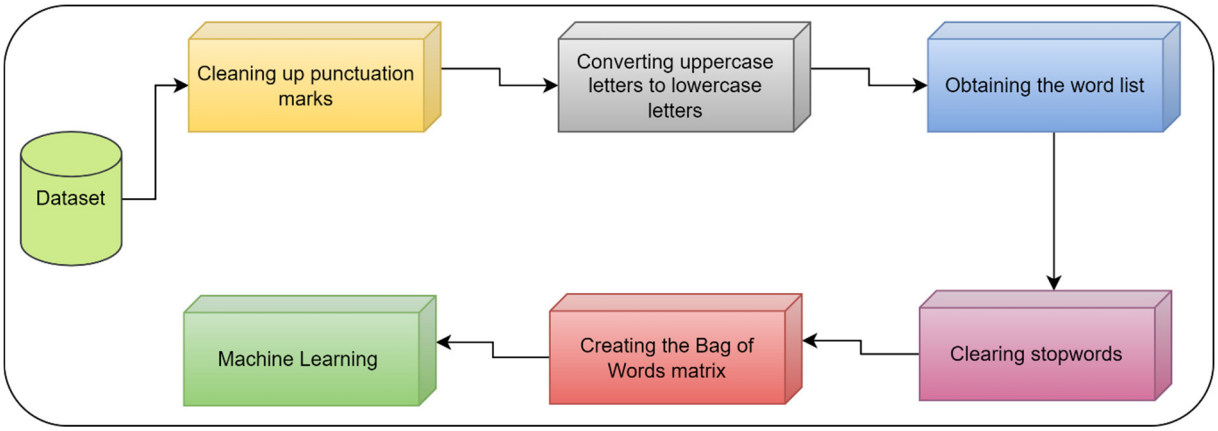

Figure 3. Data cleaning and preparation steps

In this study, before determining whether the URLs in the dataset are spam, the unstructured data in the dataset should be provided in a structured form that can be understood by machine

learning classifiers [19]. Data mining basic preprocessing should be applied to the URLs in the dataset at this stage. In these pre-processing steps, punctuation marks, Html tags, numeric expressions, and stop words are extracted from the data, and operations such as upper and lower case conversion are applied to the data. After these stages, keywords are obtained and included in the word bag called the bag of words. In this way, the most repetitive 1000 words among the words in the whole dataset are determined as keywords and a 1x1000 matrix is formed for each data in the dataset. When the entire dataset is considered complete, a 143000x1000 matrix is formed called the document term matrix (DTM). It was created by giving 1 if the keywords are included in the matrix and 0 if they are not. Thus, the dataset will be transformed into a structured form that classical machine learning classifiers can understand.

2.2. Machine Learning Techniques

Parallel to the rapid development of technology, the amount of data kept in databases also increases. For these data stored in datasets to make sense, they must be processed. These processed data can be used in different ways in different places. Health, economy, finance, agriculture, and cyber security are just a few areas. Processing and analyzing the large amount of raw data stored in databases is quite difficult with traditional database systems. Machine learning is the set of methods and algorithms necessary for processing and analyzing data. It is possible to develop a problem-specific model in machine learning.
In this study, 9 different methods accepted in the literature were used while determining the URL. In these methods, the models are first trained with the training data. Thanks to this training data, the learning process is realized. Then, when new inputs come to the trained network, the network is asked to produce the result closest to the desired value. In this way, it is aimed to place the new entry in the correct class. After the models used in the study were trained with the training data, the models were tested with the test data. Classifier and methods used in the study k-nearest neighbors(KNN) [20], Random Forest(RF) [21], Naive-Bayes (NB) [22], Gradient Boosting (GB) [23], Discriminant Analysis [24], LightGBM [25], Logistic Regression [26], XgBoost [27], Support Vector Machine (SVM) [28].

3. EXPERIMENTAL RESULTS

This study for URL classification was carried out in a Python environment. In the study, confusion matrices obtained in different models were given separately and compared. A confusion matrix is a table often used to describe the performance of a classification model on a set of test data for which the actual values are known [29]. An example confusion matrix is given in Figure 4.

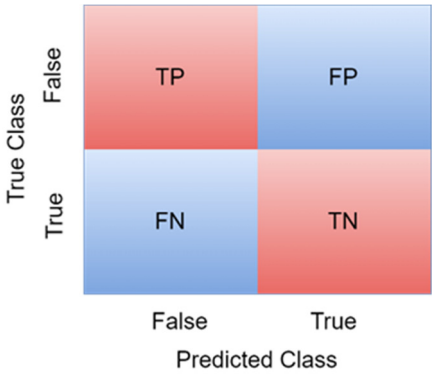

Figure 4. Confusion matrix example

In the study, 9 different parameters were used to compare the performance of the models [30]. These parameters and their formulas are given in Table 1.

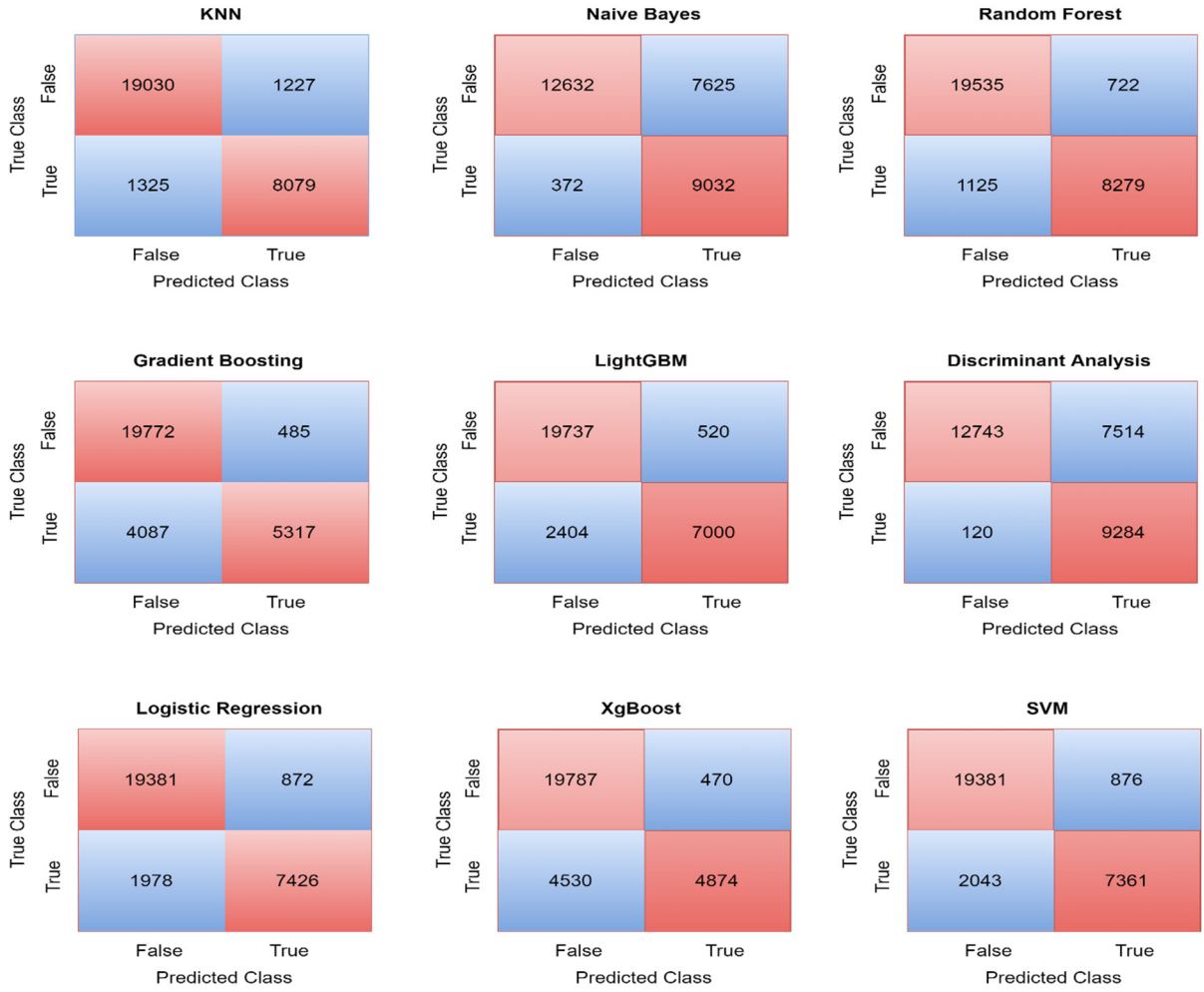

Figure 5. Confusion matrices obtained from the models

The accuracy values obtained in the machine learning methods while determining the URL are given in Table 2.

Table 1. Performance Measurement Parameters

F1-Score	**Accuracy**	**Specificity**	**Sensitivity**
F1=2TP/(2TP+FP+FN)	Acc = (TP+TN) / (Total)	Spc= TN / (FP+TN)	Sens = TP / (TP+FN)
Precision	**FPR**	**FNR**	**FDR**
PPV = TP / (TP+FP)	FPR = FP / (FP + TN)	FNR = FN / (FN+TP)	FDR= FP / (FP + TP)

In the study, to be able to classify URLs, first of all, the data in the dataset was cleaned and converted into a format that the models could understand. The Bag of Words matrices obtained in the last step of this process became the input values for the models. Confusion matrices obtained in 9 different models used in the study are given in Figure 5.

Table 2. Accuracy values(%)

KNN: 91.39	Naive Bayes: 73.03	Random Forest: **93.77**
Gradient Boosting: 84.58	LightGBM 90.14	Discriminant Analysis: 74.26
Logistic Regression: 90.39	XgBoost: 83.14	SVM: 90.15

When Table 2 is examined, it is seen that the highest accuracy value was obtained with 93.77% in the Random Forest classifier and the lowest accuracy value was obtained with the Naive Bayes classifier with 73.03%. When the confusion matrix in Figure X obtained from the Random Forest classifier is examined, it is seen that the Random Forest classifier classified 27814 of the 29661 test URLs correctly and misclassified the 1847 URLs. Other performance metrics obtained in the Random Forest classifier are given in Table 3.

Table 3. Performance metrics of Random Forest classifier(%)

F1-Score	**Accuracy**	**Specificity**	**Sensitivity**
F1= 95.49	Acc = 93.77	Spc = 91.98	Sens = 94.55
Precision	**FPR**	**FNR**	**FDR**
PPV = 96.44	FPR = 8.02	FNR = 5.45	FDR= 3.56

The proposed model is compared with similar studies in the literature in Table 4.

Table 4. Similar studies on URL classification

Study	**Year**	**Methods**	**Accuracy**
Do Xuan [12]	2020	Classifiers (SVM)	90.70%
Patgiri [13]	2019	Classifiers (SVM)	90.14%
Jain [14]	2018	SVM	90%
Joshi [15]	2019	Machine Learning	92%
Goh [16]	2015	Classifiers	85.2%-93.7%
Sun [17]	2020	Classifiers	82.9%,87.48%,91.90%
This Study	2022	Machine Learning Classifiers (RF)	93.77%

When Table 4 is examined, it is seen that the proposed model has either better or similar results than similar studies in the literature. Therefore, it is seen that the proposed model can be used in spam URL detection.

4. CONCLUSION

URL classification was made using the Bag of Word matrix in the study. Spam URLs leave users and companies in a complicated situation. These spam URLs are one of the most dangerous issues in cybersecurity. Also, spam URLs are used in fraud. To detect spam URLs, 9 different machine learning methods were used in the study. Among these methods, it has been seen that the most successful method is Random Forest. This study guides both machine learning researchers in academia and professionals and practitioners in the cyber security industry. It is among our aims to train the study with CNN, LSTM style models by using

different document matrix extraction methods. In addition, this is one of the limitations of our study.

Acknowledgments
Thank you to the researchers for sharing their datasets.

Declaration of Competing Interest
There is no conflict of interest.

Author Contribution
Muhammed Yıldırım contributed 100% at every stage of the article.

REFERENCES

[1] Adam, E.E.B., Deep learning based NLP techniques in text to speech synthesis for communication recognition. Journal of Soft Computing Paradigm (JSCP), 2020. 2(04): p. 209-215.

[2] Rajput, A., Natural language processing, sentiment analysis, and clinical analytics, in Innovation in Health Informatics. 2020, Elsevier. p. 79-97.

[3] Arthur, M.P., Automatic source code documentation using code summarization technique of NLP. Procedia Computer Science, 2020. 171: p. 2522-2531.

[4] Widyassari, A.P., et al., Review of automatic text summarization techniques & methods. Journal of King Saud University-Computer and Information Sciences, 2020.

[5] Nemes, L., A. Kiss, Social media sentiment analysis based on COVID-19. Journal of Information and Telecommunication, 2021. 5(1): p. 1-15.

[6] Neysiani, B.S., S.M. Babamir. Effect of Typos Correction on the validation performance of Duplicate Bug Reports Detection. in 10th International Conference on Information and Knowledge Technology (IKT), Tehran, Iran. 2020.

[7] Rivera-Trigueros, I., Machine translation systems and quality assessment: a systematic review. Language Resources and Evaluation, 2021: p. 1-27.

[8] Popovski, G., B.K. Seljak,, T. Eftimov, A survey of named-entity recognition methods for food information extraction. IEEE Access, 2020. 8: p. 31586-31594.

[9] Lai, C.-M., H. Shiu Jr, J. Chapman, Quantifiable Interactivity of Malicious URLs and the Social Media Ecosystem. Electronics, 2020. 9(12).

[10] Chen, Q., et al. Detecting filter list evasion with event-loop-turn granularity javascript signatures. in 2021 IEEE Symposium on Security and Privacy (SP). 2021. IEEE.

[11] Thanaki, J., Python natural language processing. 2017: Packt Publishing Ltd.

[12] Do Xuan, C., H.D. Nguyen, and T.V. Nikolaevich, Malicious URL detection based on machine learning. International Journal of Advanced Computer Science and Applications, 2020. 11(1).

[13] Patgiri, R., et al. Empirical study on malicious URL detection using machine learning. in International Conference on Distributed Computing and Internet Technology. 2019. Springer.

[14] Jain, A.K., B. Gupta, PHISH-SAFE: URL features-based phishing detection system using machine learning, in Cyber Security. 2018, Springer. p. 467-474.

[15] Joshi, A., et al., Using lexical features for malicious URL detection--a machine learning approach. arXiv preprint arXiv:1910.06277, 2019.

[16] Goh, K.L., A.K. Singh, Comprehensive literature review on machine learning structures for web spam classification. Procedia Computer Science, 2015. 70: p. 434-441.

[17] Sun, N., et al., Near real-time twitter spam detection with machine learning techniques. International Journal of Computers and Applications, 2020: p. 1-11.

[18] URL-1,https://www.kaggle.com/shivamb/spam-url-prediction, Last Accessed Date: 01.01.2022.

[19] Bingol, H., B. Alatas. Rumor Detection in Social Media using machine learning methods. in 2019 1st International Informatics and Software Engineering Conference (UBMYK). 2019. IEEE.

[20] Zhang, M.-L., Z.-H. Zhou, ML-KNN: A lazy learning approach to multi-label learning. Pattern recognition, 2007. 40(7): p. 2038-2048.

[21] Pal, M., Random forest classifier for remote sensing classification. International journal of remote sensing, 2005. 26(1): p. 217-222.

[22] Rish, I. An empirical study of the naive Bayes classifier. in IJCAI 2001 workshop on empirical methods in artificial intelligence. 2001.

[23] Friedman, J.H., Stochastic gradient boosting. Computational statistics & data analysis, 2002. 38(4): p. 367-378.

[24] Klecka, W.R., G.R. Iversen, and W.R. Klecka, Discriminant analysis. Vol. 19. 1980: Sage.

[25] Ke, G., et al., Lightgbm: A highly efficient gradient boosting decision tree. Advances in neural information processing systems, 2017. 30.

[26] Wasserman, S., P. Pattison, Logit models and logistic regressions for social networks: I. An introduction to Markov graphs andp. Psychometrika, 1996. 61(3): p. 401-425.

[27] Chen, T., C. Guestrin. Xgboost: A scalable tree boosting system. in Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining. 2016.

[28] Suykens, J.A., J. Vandewalle, Least squares support vector machine classifiers. Neural processing letters, 1999. 9(3): p. 293-300.

[29] Eroglu, Y., et al., Diagnosis and grading of vesicoureteral reflux on voiding cystourethrography images in children using a deep hybrid model. Computer Methods and Programs in Biomedicine, 2021. 210: p. 106369.

[30] Yildirim, M., A. Çinar,, E. Cengİl. Classification of flower species using CNN models, Subspace Discriminant, and NCA. in 2021 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT). 2021. IEEE.