World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Developing data mining algorithms for predicting cyber security risks using predictive analytics

Zahraa Raji Al-zobaiby *

Department of invitation and thought Al-imam Aladham college, Baghdad, Iraq.

World Journal of Advanced Research and Reviews, 2025, 25(02), 585-592

Publication history: Received on 23 December 2024; revised on 02 February 2025; accepted on 05 February 2025

Article DOI: https://doi.org/10.30574/wjarr.2025.25.2.0344

Abstract

Data mining methods used as one of successfully potential solution against cyber risks, big data and electronic thread increasing faster in the last year's, so it forms a challenge to cyber security, data prediction used as one of the fundamental tools to predict cyber risks and improving security methods.

Cyber security is one of important popular challenges in the current era where cyber-attacks and risks increasing, so fast it is very important to develop tools and techniques of cyber security, data mining techniques one of the important method used to solve cyber security problems. In this study we took support vector machine with random forest as prediction of risks tool then tested on a set of chosen data to detect attacks and risks on information and demonstrate the results, merging and make a comparison between them to gain best way for predicting of cyber security risks and determine the unusual data patterns that consider suspicious activity and improving the response to them.

Keywords: Data Mining; Cyber Security; Predictive Analysis; Random Forest and Support Vector Machine

1. Introduction

Rapid growth of technology in computers and communications consequences a variety of cyber-attacks that repeated while many weak point still found in systems and applications , data mining one of different ways trying to investigate and reduction these electronic attacks risks. Data miming is the process of extracting useful information from huge amount of data according to specific patterns, but it is not having important or improved of their technology in cybersecurity [1].

In this paper we take a cybersecurity as a comprehensive in the first place for protect computers from thieving, damaging and illegal activity which one of the big challenges in the meantime that is needs very specialized methods in electronic system.

Data mining technologies aims discovering hidden patterns and developing a predictive model's that is active in information process security challenges using classification, temporal analysis and many other discovering techniques have been developed as strong solutions to deal with the recent attacks. For data mining the data is very important because the model is learned from the given data , it is very necessary to have understanding the data and its component to study and analyze it, it offers systematic models discovering and detection of threats and attacks, monitoring patterns and behavior using data mining algorithms [2].

The fundamental issue of examination in this paper is how to data mining strategies discovering cybersecurity attacks and risks, harmful programs, threat detection and malware discovery , cybersecurity is a set of techniques designed to

* Corresponding author: Zahraa Raji Al-zobaiby.

Copyright © 2025 Author(s) retain the copyright of this article. This article is published under the terms of the Creative Commons Attribution Liscense 4.0.

protect computer devices , organizations , developers and data against illegal access and attacks. Every systems have security system symbolize by firewall , antivirus application, intrusions detection programs, attacks on systems may be outsider or inner in Figure (1) shows cybersecurity variables [3].

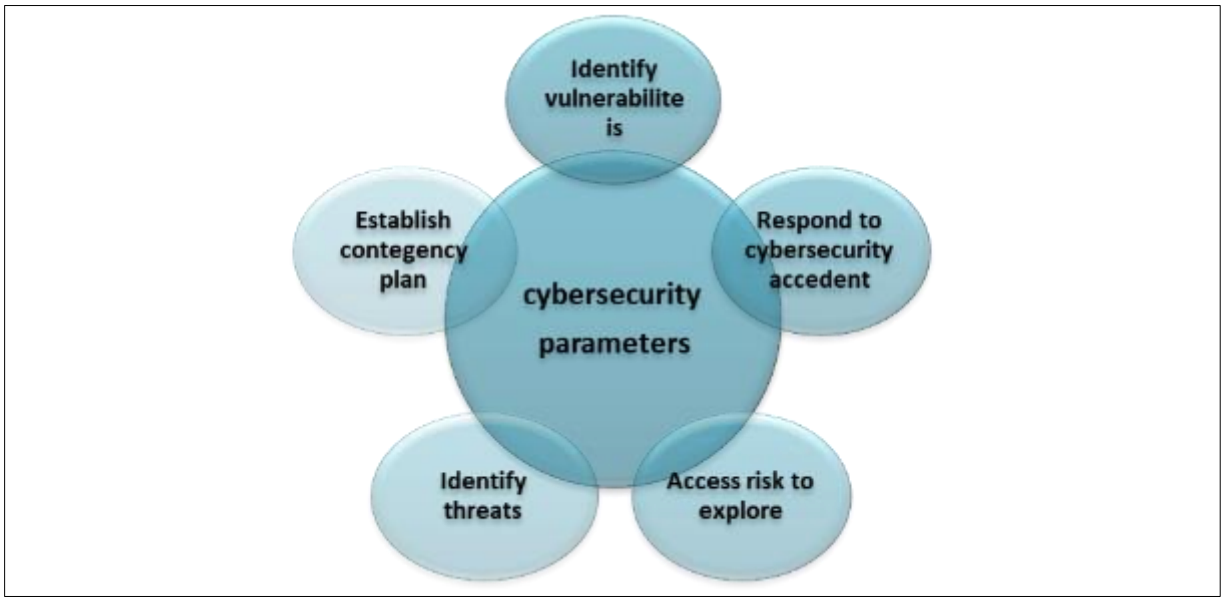

Figure 1 cybersecurity parameters

2. Literature survey

Security of data and information has been basic factor to determine successful from the starting point . data and information was collected manually at the past and had boundaries access to expert for saving data . The rapid growth of technology appears the need of information preserving so security becomes a real problem as result for this growths risks , this kind of security arise in environment of computer, network or cloud.

According to the conclusion of literature review data mining is on of basic elements to handle challenges belongs to cybersecurity information . In [4] data mining presents different types of techniques that can be used to assist determining and avoiding of possible cybersecurity risk one of the used method is deep stacking network woks basically on analyzing data and it's relation . The author of [5] presents a review of methods of data mining attacks detection and many types of intrusion detection based on network and host as an example it is bounded some applications like suspicious site classification, anomaly detection and reuse attack reduction.

A method of intrusion detection as un ordinary detection is one of a significant kind points of interest in [6]. In [2] presents a study about extraction of data and artificial intelligence tools to improve evaluation of attack discovered and the importance of data mining in the future security . In [7]dealing with software security as an efficient security development technology to reveal software attack on private and sensitive people data where the main aim of them is to harm people and systems or the whole world , because the attack is dynamic and it is very difficult to reveal patterns manually for that automated supervised machine learning used for fraud and virus detection as effectively recently.

3. Data mining for Cybersecurity

Cybersecurity" is a set of rules and technologies that are together used to protect computers, systems, networks, and data against unregistered access and attacks". In other words making efforts to protect the confidentiality of information management systems from possible attack by applying a different of cyber defensive systems. These efforts have led to the from record files and efficacy on the network to detect and bounding of intrusion cases. The process of examining data from a particular source and assembling that input into useful information is referred to as data mining, it is also known as knowledge discovery of protecting online information Mining methods used to determine potential situation risk [3,8].

Information protection through cryptography by converting data into a format that is not readable (cipher text). Only experts that having the hidden key can interpret this cipher text. In the present people and users, also organizations, governments, educational institutions, our financial information and our enterprises, see cybersecurity as an important part of the technology [8].

In our research we develop using of data mining techniques in cybersecurity applications to detect unusual patterns or behaviors and analyze it to trace suspicious and its creator . Classification can be used to collect various internet attacks using its personal information and detect attacks when it appears, also using predictions to detect future potential attacks depends on the information gained by attacker.

Applying data mining for cybersecurity pros by [9]:

- Useful insight from existing data
- Identification of security flaws and blind
- Detection of zero day attacks
- Detection of intricate and masked attack patterns

While it has some cons as:

- Need for deep data science expertise
- Time and effort to prepare database for mining
- Constant effort on updating classifiers and mining techniques
- Risk of disclosing sensitive information from database
- Manual verification of data mining results

Data mining basic techniques has both predictive and descriptive depending on the data analysis method shown in Figure (2) below [10].

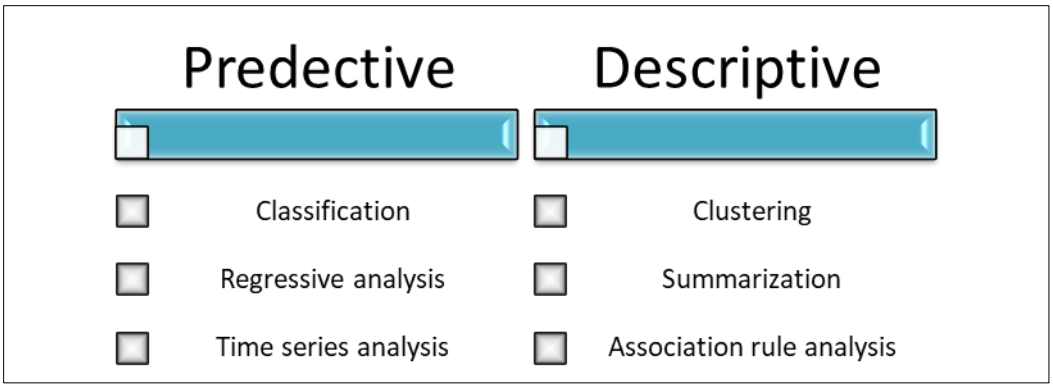

Figure 2 Data mining techniques

4. Data collection

Collecting data is basic step in data mining to predict cyber risks choosing data carefully, Our used data are automatically downloaded from the internet websites Kaggle Dataset (https://www.kaggle.com) that contain a set of cyber data like records of cyber-attacks and network

4.1. Data sources

All data sources text in English only because mixing with another language is not preferred for the used algorithm. These data from posts, internal records , firewalls and threat intelligence platform.

4.2. Data collection tools:

Using snort free and open source tool for collect and analyze data intrusion detection, it also used for discovering security threats , analyze net traffic, block attacks and construct detailed reports depend on programming grammar for suspicious patterns of traffic it is an efficient tool easy to merge with data analyzing tools [11].

4.3. Data manipulation steps:

- Data cleaning: an operation used to ensure quality of used data that train and analyze algorithms to predict cyber risks and remove odd and missed value data cleaning steps as follows:
  - Remove missed and illegal value: determine fields that contains missed value like traffic records not contain address or target and remove them if it was high percent.
  - Remove massive like ordinary frequent activity: determine frequent records like frequent data in logs then remove the frequency to obtain analysis accuracy [12].
- Features selection: it is an essential step in data manipulation specially in cybersecurity by determining most important features to improve model performance , reducing data size , reduce training time and improve interpretation ability. In this paper we use random forest technology that set feature organization depends on the importance , feature selection challenges in cybersecurity symbolize as dealing with unbalanced data, data noise, temporal data and huge data .
- Dealing with unbalance data: a set of data contain huge not stabile frequent group of data it is famous in many cybersecurity applications where attacks rare. Most used algorithms assumed the data balanced but dealing with unbalanced data needs efficient algorithm, Applying techniques like:
  - Over sampling: using smote (synthetic minority over-sampling technique) algorithm by adding new examples to the less common group depends on recent data value ,trained datasets from Kaggle to the implanted smote algorithm in python .
  - Under sampling: reduce extra end ordinary activities to be balanced with the rare group , this type has benefits like reduce data size and fast training while its defects losing important information is possible Figure (3) illustrate them [13].

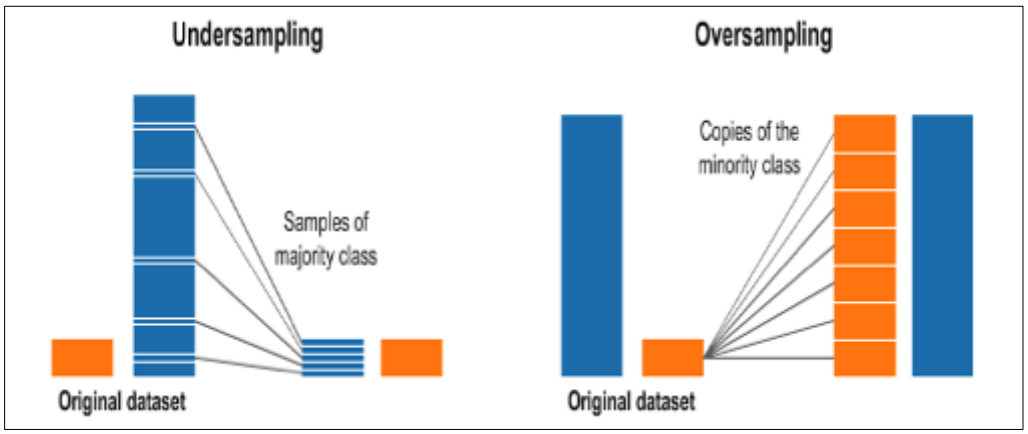

Figure 3 oversampling and undersampling [5].

5. Support vector machine (SVM) and Random forest methods

Support vector machine is one of supervised data mining classification algorithm (supervised learning), basically used for regression and classification depends on finding a hyberplane interval between different datasets ,Basic concept of SVM technique are decision boundary, support vectors and margins where boundaries separate different sets and find the interval that achieves greater space with closer point from every set to ensure best performance dealing with new data, support vector is the interval closer point used to determine best interval ,is most important to set final interval and margin searching for largest possible margin where SVM is the space between the interval and closer point for every set to achieve the best margin [9].

SVM workflow first entered data separated into train and test by (30%-70%) as features (train set) each had characteristics linked by label, secondly choosing a break by finding two dimensions linearly separated if the data has breakable capability in higher level separate between different levels, third greatest margin through trying to find greatest margin possible break between different SVM group , finally dealing with linearly non-breakable data by converting data to kernel trick if it linearly non-breakable data using special techniques named high dimensional space though it can be linearly breakable [14].

Kernel in SVM have many types as :

- Linear kernel: used when the data capable to separate.
- Polynomial kernel: used to separate data contained multidimensional relation .
- Radial basis functions kernel (RBF): is the most widely used to separate complicated nonlinear data.
- Sigmoid kernel: used in special case

Preparing data by splitting into two sets train and test , choosing proper kernel depends on data nature to choose the kernel then train the model using train set and finding the proper break last test the model using test set Figure (4) illustrate SVM algorithm steps [7].

Random forest is a very popular machine learning algorithm used in classification purposes depends on decision making on a set of trees this technique one of the ensample learning methods works as committee merging multi decisions to improve accuracy, reducing partiality and contrast through applying different data trees so the result depends on many trees not only one will reduce the contrast . The main idea of random forest depends on building many decision trees each one builds using random part of data Bootstrap Sampling techniques to predict then merging all commonly trees outputs [14].

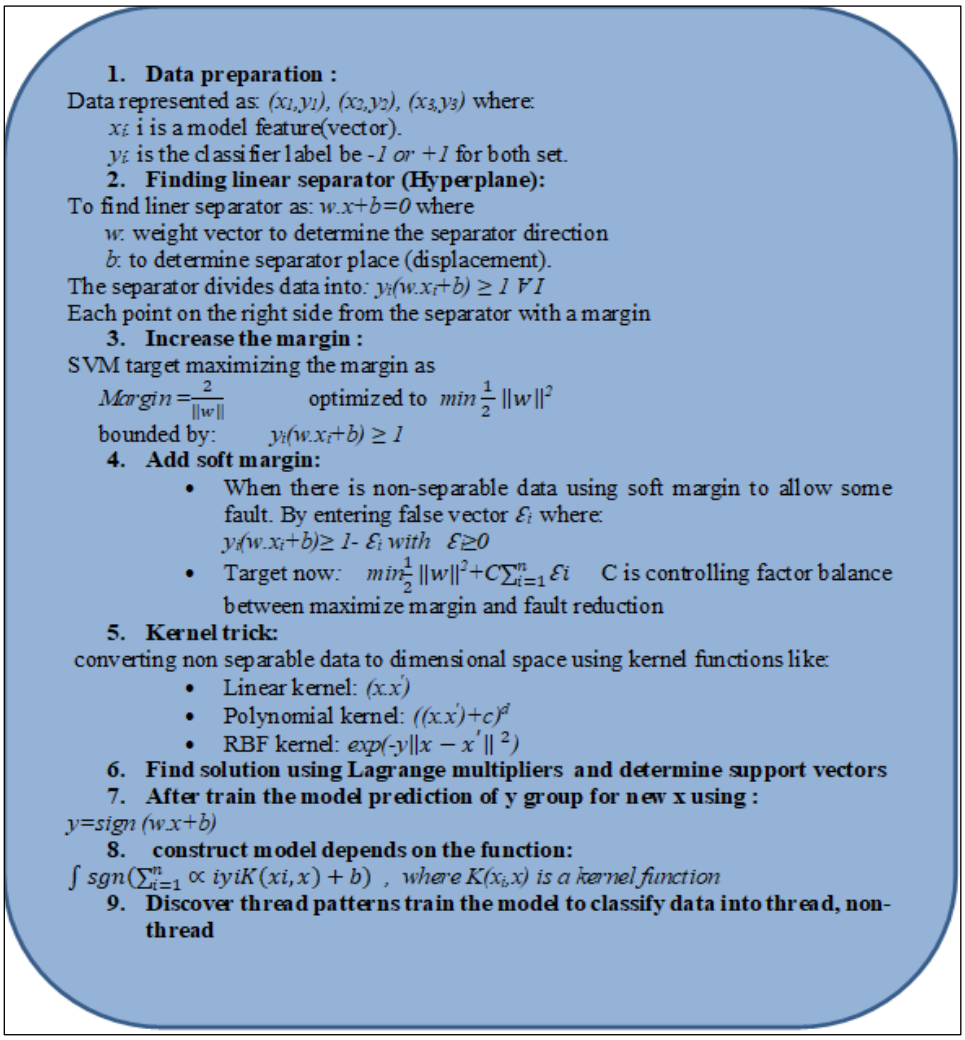

Figure 4 Developed SVM Algorithm [16]

Main specifications of random forest are:

- **High accuracy** : so strong performance even with very complex data.
- **Overfitting resistance**: uses of multi trees that avoid depending on one tree.

- **Ability of applicants**: on classification and regression .
- **Very active with missing data**: provides predictions even when some data missed.
- **Applicant**: with high dimensional data.

Random forest also has some defects are:

- **Slow ratio** : be slow with huge data.
- **Interpretability**: it is hard to interpret final model because of multiple result tree.
- **Need to set estimators** : like number and depth of the trees

Continuous improvement of model by updating and repeat training with new available data , adaptability to new threads by merging online learning techniques to update the model dynamic [15].

6. Results and evaluation

After applying a model on training data, the developed model can be saved so it can be used in various systems later. The evaluation classification metric used for DM as shown in Table (1) binary confusion matrix where TP, TN, FP and FN represent True Positive, True Negative, False Positive and False Negative.

Table 1 Binary Confusion Matrix

	Actual class: X	**Actual class: not X**
Predicted class: X	TP	FP
Predicted class: not X	FN	TN

The used metric in supervised learning problems are as follows

- **Accuracy correct:** calculate as (TP+TN)/ (TP+TN+FP+FN). Good measures achieved with Balanced classes.
- **Positive predictive value (PPV) or precision:** calculate as TP/(TP+FP). Items ratio categorized correctly as X to all items categorized as X.
- **Recall or probability of detection PD:** calculate as TP/(TP+FN) items ratio classified correctly as negative (not X).
- **FP rate or fall-out(F1 score):** calculate as FP/(TN+FP) [1,16].

Suggested model using SVM and random forest algorithms merging them to detect cyber security attacks like DDos attack or data thief and comparing the result depending on the evaluation metrics table 2 shows the results of evaluation algorithms

Table 2 resulats and evaluation

Algorithm	**Accuracy**	**Precision**	**Recall**	**F1 score**
SVM	88%	86%	84%	85%
Random forest	91%	89%	88%	88.5%
Merging	93%	91%	90%	90.5%

Note that SVM is accurate in small thread detection but slower when dealing with big data, while random forest results higher performance in pattern recognition may be contain some noise, merging them enhancing predicting performance through combining accuracy of SVM with fasting of random forest. Figure (5) illustrate the SVM, random forest and merging them detection cyber security performance.

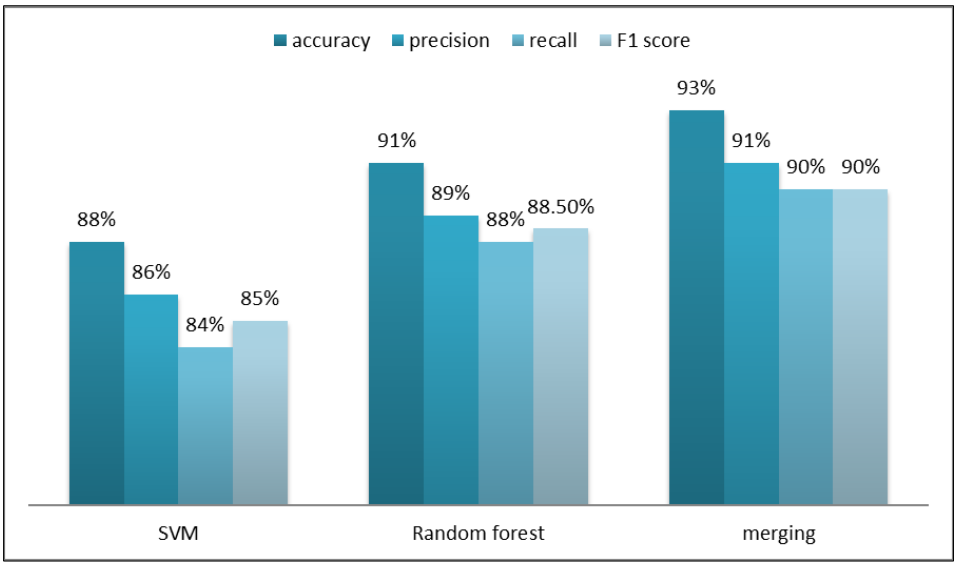

Figure 5 SVM, random forest and merging them detection cyber security performance

7. Conclusion

Developed algorithm used to contributes in revealing threads patterns electrical using advanced SVM with predictive analysis that helps in cyber security strategies in modern systems. The system use SVM , random forest algorithm and merging them to detect cyber security attacks and reducing their risks . As shown in the above evaluation results the SVM is slower with big data but more accurate than random forest while random forest is faster than SVM but not accurate as random forest , merging them more efficient in cyber risk detection by gaining more accuracy and faster investigated results.

There are some anomaly characteristics of cyber data problems making data mining techniques difficult to use these are necessity of repeating model training, improving this for more investigate and research through daily updating of the used model for training to detect misuse and cyber risk attack as one of suggested things to give more efficiently results.

References

[1] Deepa D. Shankar, et al. Data mining for cyber biosecurity risk management – A comprehensive review. Computers & Security. 2024, pp. 1-13.

[2] Aithal, P. S. "Data Mining and Machine Learning Techniques for Cyber Security Intrusion Detection," . 2020.

[3] al., Israa Ezzat Salem et. introduction to data mining techniques in cybersecurity. Mesopotamian Journal of Cybersecurity . 2022, pp. 28–37 .

[4] Kong J., Yang C., Wang J., Wang X., Zuo M., et al.,. "Deep-Stacking Network Approach by Multisource Data Mining for Hazardous Risk Identification in IoT-Based Intelligent Food Management Systems,". Computational Intelligence and Neuroscience. 2021, pp. 1-16.

[5] P., Patond K. and Deshmukh. "Survey on Data Mining Techniques for Intrusion Detection System,". International Journal of Research Studies in Science, Engineering and Technology. April 2014, pp. 93-97.

[6] Nieves J. F and Jiao Y, C. "Data clustering for anomaly detection in network intrusion detection,". Research Alliance in Math and Science. August 2009, pp. 1-2.

[7] al., M. R. Ul Islam et. "Automatic detection of NoSQL injection using supervised learning,". IEEE 43rd . 2019.

[8] Bavani M., Kevin W. and Mahedy M. Data mining for security applications . IEEE/IFIP International Conference on Embedded and Ubiquitous Computing. 2014, pp. 585-589.

[9] Anna L. Buczak, Member, IEEE, and Erhan Guven. ASurvey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection. IEEE COMMUNICATIONSSURVEYS&TUTORIALS,VOL.18,NO.2. 2016, pp. 1153-1176.

[10] Varsha P.Desai, K.S.Oza and P.G.Naik. Data Mining Approach for Cyber Security. International Journal of Computer Applications Technology and Research, Volume 10–Issue 01. 2021, pp. 35-41.

[11] Maaike H. T. de Boer 1, Babette J. and et al. Text Mining in Cybersecurity: Exploring Threats and Opportunities. Multimodal Technol. Interact. ,vol 3. 2019.

[12] Data Mining: Dirty Data and Data Cleaning. Dwivedi, Santosh Kumar Singha and . Rajiv Kumar. India : s.n., 2020. International conference on Recent Trends in Artificial Intelligence, IOT, Smart Cities & Applications . p. 5.

[13] Zhang, X. Research on Data Cleaning in Data Mining. Journal of Theory and Practice of Engineering Science, 5(1). 2025, pp. 26–32.

[14] Bedi, Mayank Arya Chandra & S. S. Survey on SVM and their application in image classification. International Journal of Information Technology ,vol. 13. 2021, pp. 1-11.

[15] Paul, Angshuman and al., Dipti P M and et. Improved Random Forest for Classification. IEEE Transactions on Image Processing, vol. 27, issue:8. 2024, pp. 4012-4024.

[16] S., Derek A. Pisner and David M. support vector machine . Machine Learning methods and applications to brain disorder . s.l. : Elsevier , 2020, pp. 101-121.