Library Progress International *Print version ISSN 0970 1052*
Vol.44 No.3, Jul-Dec 2024: P.7576-7583 *Online version ISSN 2320 317X*

Original Article *Available online at www.bpasjournals.com*

Support Vector Machine based Pattern Taxonomy Classification Method for Web Usage Mining

G. Punithavathi[1], Dr. R. Sankarasubramanian[2]

[1]Ph.D., Research Scholar, PG and Research Department of Computer Science, Erode Arts and Science College, Erode-9, Tamilnadu,

[2]Principal, Erode Arts and Science College, Erode-9, Tamilnadu,

How to cite this article: G. Punithavathi, R. Sankarasubramanian (2024) Support Vector Machine based Pattern Taxonomy Classification Method for Web Usage Mining. *Library Progress International,* 44(3), 7576-7583.

ABSTRACT

One of the most active research topics is machine-learning-based text classification, which has several uses, such as topic modelling, rating summarisation, reviews, hate speech identification, spam detection, and sentiment analysis. There are differences in the datasets, training techniques, performance assessment techniques, and comparison methods employed in commonly utilised machine-learning-based research. In this study, we conducted a survey of 224 studies that used machine learning for text classification that were published between 2003 and 2022. However, for complex typesetting articles, rule-based solutions come with a large coding cost. However, the mere use of machine learning techniques necessitates the expensive annotation of complex content types inside the document. Moreover, relying exclusively on machine learning may result in instances when patterns that are readily identified by rule-based techniques are inadvertently retrieved.To scan text, photos, and HTML documents and return results to the search engine, web content mining technologies were required. It directs the search engine to deliver more fruitful outcomes for each query according to its significance. The paper the proposed method WBSVM is analysed different web content mining tools for the extraction of relevant information from the corresponding web page.

Keywords: Machine Learning, Web Crawling, Web Usage Mining, Pattern Taxonomy, Pattern Discovering, Web Based Support Vector Machine

INTRODUCTION

Web usage mining involves searching for and analysing patterns, trends as well as other pertinent facts in the huge amounts of data that utilizers generate while generating and interacting with websites. This belongs to the broader category of data mining that tries to create utlizeful information as well as insights from sizable datasets. The websites are the major medium for exchanging information, communicating as well as doing a different possibilities of transactions on the internet, as well as doing a diversity of the transactions on the internet that has become an essential section of our day to day life. Web server logs stay tracking the utilizer activites every time they try to access a website what they search, what they activated in a regular search including userful information such as page views, clicks, navigation paths, timestamps and many more. The usage of website they try to make its widespread as well as interactive medium to disseminate information. Data can be extracted from different sources utilizing the web that could be utilized in different researches.

Data mining is a process that helps to extracting the information from a given data set to identify the trends, patterns and utilizeful data. Web usage mining is one type of data mining techniques that handles the combination of structured and unstructured data. When viewed in terms of data mining, web mining is intended to cluster, associate and analyse information from web data sources.

Web mining has gained an important role in the data mining sector due to the exponential growth of web data as well as the growing necessary for a more sensible as well as logical search system. It is evident that a web mining framework will be useful once it is able to respond quickly and accurately to the needs of the users. The process of mining useful patterns from a web database of a large scale takes a long time. In addition to find the limited amount of input data that is connected to a particular device or utilizer community.

Initially filtering the most irrelevant data quickly processing the filtered data as well as then returning the most important documents that can fit the necessary of the utilizers is appropriate. Web usage mining focuses on the exploration of utilizers behaviour information when surfing websites as well as web applications. A method of utilizing data mining techniques as well as algorithms by extracting data from web documents as well as facilities, web content, hyperlinks as well as server logs to collect information from the web. In order to gain insight into trends and the industry as well as the utilizers in general, the purpose of web mining is to search for patterns in web data by collecting as well as analysing content. Web mining is a looping method in which prototyping plays an important role in order to quickly experiments with various alternatives as well as to integrate the information gained during the process itself during existing looping methods. By utilizing the variety of data mining techniques, machine learning algorithms as well as statistical analysis. The web usage mining aims to make sense of this unstructured data by spotting usage trends as well as utilizer behaviours.

II. RELATED WORKS

From the literature review, we found that in the web mining clarification of an event, logs that could often be delivered with a combination of concept that logs often be delivered with a combination of concepts, event/trace sorting like as based on attributes as well as clustering of the trace. Researchers have mainly focused on the work going in the field of process mining.

According to Roy et.al [15], web mining is the technique of identifying similar pattern as well as information inside a weblog file. This is made up of pages in different format like as pictures, HTML documents etc,. Additionally, the record's data is still growing at an extremely rapid rate. This result extracting information from it is an extremely challenging and undertaking the concepts. In order to retrieve the necessary data from the web page. Web log file mining is needed for analysing the data. The researcher reviewed earlier research as well as topics that have been studied in web prospecting. Describe the principles of web mining. The utilizers learn how to use the web usage mining techniques as well as algorithms to acquire information as well as observe visitor patterns.

According to Bharathi et.al [14], the world wide web's constant expansion has resulted in long access delays. To lesson this prefetching techniques have been to employed to forecast utilizers browsing behaviour as well as retrieve the web pages before they are explicitly requested. Researchers have long struggled to create near perfect predictions about consumers search behaviour. Several web mining techniques were utilized to accomplish this. However, it is clear that each of the strategies has its own set of disadvantages.

A novel strategy has been developed for developing a hybrid prediction model that combines utilize mining as well as content mining techniques to address the individual limitations of both approaches. In order to improve web page prediction and they suggested method employs N-gram parsing in conjunction with query click counts to acquire more contextual information. The suggested hybrid strategy was evaluated utilizing AOL search records, as well as the results reveal a 26% increase in prediction precision are increase in hit ratio on average when compared to other mining techniques. This paper were examined the greater methods as well as algorithms are utilized. The transition system as well as region basis analysis were performed utilizing deterministic algorithms that is always create repeatable models. All the data as well as web mining is persistent for the assumed variable input.

Asadianfam et.al [13] propose a novel technique to web usage mining depends on case based reasoning. The case based reasoning approaches are a knowledge based problem solving method that relies on the reuse of prior data handling methods. This can be utilized as an effective guide for tackling novel difficulties. The web personalization solutions that can customize the next batch of visited pages to individual utilizers based on their interests as well as navigational behaviours. The suggested architecture is made up of several components, including basic log pre-processing, pattern finding methods case based reasoning as well as peer to peer similarity clustering association rules mining approaches as well as recommendations.

Sharma et al [10] is the exponential development of Internet utilizers as well as traffic, information searchers rely heavily on search engines to get relevant information. Because of the widespread availability of textual, audio, video as well as other types of content, search engines responsibilities have developed. The search engine offers appropriate information to Internet utilizers based on their query, such as content, link structure, as well as so on. However, it does not ensure the accuracy of the information has to be ranking module plays a critical role in a search engine's concert. The presentation of the ranking module is helps to find the link structure of web pages (WSM) as well as their content (WCM). Web mining is most important for determining the ranking the web pages.

III.PROPOSED WORK

The proposed works determine that all the documents are divided into paragraphs. So that a specific document is divided in to the collection of paragraphs PS(d). Let D be a training set of documents, that consists of a set of positive documents, D+ as well as a set of negative documents, that consists of a set of positive documents. D+ and a set of let

T={t1,t2,…tm} be a set of terms which can be extracted from the set of positive documents, D+. In normally there are two phases. Training and testing on this training phase the d- pattern in positive documents (D+) depends on a minimum support are found as well as evaluated term supports by deploying d- patterns to terms. During the testing phase, there is proposal to adjust the support terms by incorporating noise from negative documents in D utilizing an experimental coefficient. Once the incoming documents are received they can then be organized according to these weights. The suggested method aims to refine the precision for accessing term weights because of identifying the patterns are more precious than entire documents and also suggested the addressing the drawbacks of the phrase based approach by utilizing the pattern based approach. Mining techniques for patterns can be employed to identify various text patterns. The following are some important definitions:

1. **Frequent and closed Patterns**

   A term set X is considered to be a frequent pattern if its supports exceeds the minimum support denoted as min_sup otherwise min_sup a. A pattern X that is also a termset is classified as closed if as well as only if X equals Cls (X).

2. **Pattern Taxonomy**

   By employing is a relation pattern can be organized into a taxonomy. The semantic statisticsinside the pattern taxonomy is utilized to augment the presentation of utilizing the sealed pattern in text mining.

3. **Closed Sequential Patterns**

   A sequential pattern X is labelled as a regular pattern if its relative support or absolute support exceeds minimum support, known as min_sup. A frequent sequential pattern X is deemed closed if there are no super pattern X of X such that Sup (X) – sup(X).

4. **Pattern Taxonomy Process**

The pre-processed documents are divided into paragraphs, resulting in a set of paragraphs PS(d) for each document d. Let D represent a training set of documents, comprising a set of positive documents, D+, and a set of negative documents, D-.

5. **Pattern Deploying**

The pre-processed documents are divided into paragraphs, resulting in a set of paragraphs PS (d) for each document d. Let D represent a training set of documents, comprising a set of positive documents, D+ and a set of negative documents, D-. From the set of positive documents, D+, a set of terms can be extracted. The frequent pattern as well as the method of interpreting discovered patterns that involve summarizing them as d-pattern to precisely assess term weights. This method is very useful to decrease the side effects of noisy pattern of the low frequency problems.

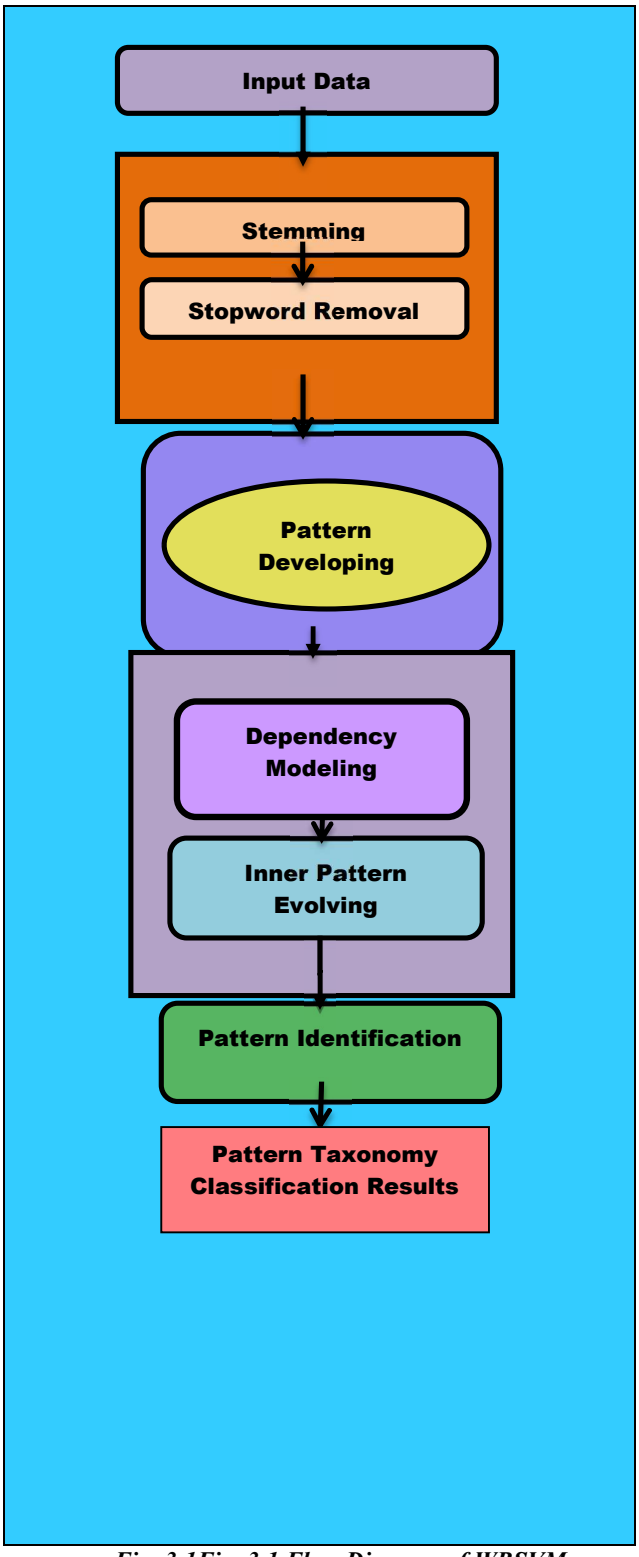

Fig. 3.1Fig. 3.1 Flow Diagram of WBSVM

The utilizer represent the database as well as from the utilizers are recommended to their interested browsing prospect that meets the necessity of the explicit utilizer at a specific time.

Algorithm of WBSVM

Step 1:	***Load the important libraries. import pandas as pd***
Step 2:	***df = pd.read_csv("mydataset.csv")***
Step 3:	***For each (iSn)***
Step 4:	***Y = df[['Var_Y']]***
Step 5:	***svm_clf = svm.SVC(kernel = 'linear')***
Step 6:	***y_pred_test = svm_clf.predict(X_test)***
Step 7:	***metrics.accuracy(y_test, y_pred_test)***
Step 8:	***df<- read.csv("mydataset.csv")***
Step 9:	***Itemp = Itemp {i}***
Step 10:	***For each (iItemp)***
	train <- df[samp,], test <- df[-samp,]
Step 11:	***accuracy(test$Var_Y, y_pred_test)***
Step 12:	***For each valid α do Call CloSpan(s α, Ds α, minsu L***
Step 13:	***End***

3.1 DataPreprocessing

For the data set as we utilize the Google web API to collect the document to be collected 64 documents on ipad concept it has 9998 features.

StopWord Removal

Stop words are language specific functional words, are frequent words that carry no information (i.e pronouns, prepositions, conjunctions). In English languge there are about 400-500 stopwords. In sample the words include "the ", "of", "and ", "to". The initial step docing the pre-processing is to neglect these stopwords that has to proven as very important.

Stemming

Stemming procedures are utilized to determine the root or stem of a word. Stemming reduces words to their stems, utilizing a large amount of language dependent linguistic knowledge. The premise behind stemming is that words with the same stem or word root typically convey the same or similarly close concepts in literature allowing the words to be confused by employing stems. The words user, users, used and using can all be shortened to the word "USE". The porter stemmer method, the most widely utilized in English words is utilized in the web search engine documents. Over stemming occurs when two words with distinct stems are stemmed to the same root.

3.2 Pattern Discovering

Data mining techniques such as association rules, sequential pattern discovery, clustering and classification were employed by recommendation systems to develop as well as offer recommendations based on utilizer behaviour as well as attributes. Simply derive recommender systems attempt to predict the utilizer's style of thinking, determine the best as well as closet interests, as well as recommend the choices to the user. The CBR approach is notable for modelling human behaviour while coping with novel solutions. As a results earlier problem solving experience are utilized to guide the solution to the new difficulties.

The WBSVM approach contains four stages such as retrieval, reuse, revision and retention. Checking the retrieval correctness of the cases means that if one case is called from the case database, the system should deliver the information. Several cases from the collection were chosen as well as tested for this purpose. The test finding reveal that the average value as well as find the similarity between the experimental cases in the case of database is 100%. As a result it is possible to determine that the system recovery's accuracy is appropriate for bidding on the utilizers navigation web pages.

3.3 Pattern Deploying

Deployment patterns are ways for adding new features or upgrades to an application in a controlled as well as organized manner. The purpose is to reduce downtime, verify that the new features are stable as well as work effectively as well as occasionally allow for testing with a limited sample of utilizers.

3.4 Pattern Taxonomy Classification

By automatically evaluating repair verbatims, the bi level feature extraction based text mining for defect diagnosis solves the aforementioned problems. Our core methodology is to extract syntactic as well as semantic fault features that are merged to provide the intended results. The proposed feature fusion of two levels might increase the accuracy of fault identification for all fault classes, especially minority ones, because the extracted features at each level has constraints as well as focus only on a particular component of feature spaces.

A popular supervised learning technique for building different machine learning models that aim to predict class labelled data in a linked dataset is classification. The machine learning literature has presented a number of

classification methods for a variety of uses. The support vector machine, random forest, decision trees, neural networks and naïve bays are the most well know classification techniques. Moreover the web based support vector machine approach is a potent method for classification that might be applied to regression issues. This means that unless the training dataset comes across a specific test question for prediction utilizing this method, it won't be completely processed. In reality, an extensive search in large dimensions can be carried out via WBSVM and the training set is conceptualized as an m-dimensional space of patterns for finding the K-closest tuples to a test query in the K-Nearest neighbours classifiers.

IV. RESULTS AND DISCUSSIONS

The ability of an information retrieval system to return pertinent documents, as well as the accuracy as well as Mean Squared Error of these retrieved documents, is commonly measured.

Mean Squared Error=

$$\frac{|relevant\ documents\ \cap retieved\ documents|}{|retrieved\ documents|}$$

---- (1)

Reuters-21578 and kaggle are the dataset supports a variety of web data publication formats, as well as encourage dataset publishers to share their data in an accessible, non proprietary format if possible. This should not be applied not only are open, manageable data format helps to improved, maintained on the platform, as they are similarly easier to work with for more people irrespective of their tools. Dividing the data into two sets, utilizing one for model training as well as the other for model testing. Mean Squared Error parameter values are calculated utilizing the formula in Eq(1). In equation (2) calculated the cross validation as well as find the real world scenarios as well as testing the out of sample webpage. To obtain a more accurate estimate of the model's performance, this can be repeated several times with various divisions. A model validation approach called cross validation is utilized to determine how well the data mining analysis findings will transfer to a different collection of data. This is generally applied in a situations where predicting outcomes is the major objective as well as one wishes to gauge how well a predictive model would work in real world scenarios. Another name for cross validation is out of sample testing.

$$cv_{(k)} = \frac{1}{k}\sum_{i=1}^{k} MSE_i$$

-------(2)

These binary measures benefit to compute additional information retrieval metrics which is F-measure.

$$F - measure = \frac{2*precision*recall}{Precision+reca}$$

-------(3)

Accuracy is utilized as a statistical measure of how well a binary classification test correctly identifies or excludes a condition is

$$Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$$ ------*(4)*

The accuracy of system is calculated utilizing the cross validation in this method as well as calculate the values utilizing given formula. The results obtained by the system in first investigates. Locating the related files is without problems utilizing the keyword problems.

Table. 4.1 Enhanced WBSVM Classification Method

Methods	MSE (%)	CV (%)	F-Measur (%)	Accuracy (%)
SVM	**89.24**	**87.45**	**88.23**	**89.54**
PDM	**90.12**	**90.34**	**87.59**	**91.67**
WBSVM	**90.23**	**90.16**	**92.67**	**95.12**

In table 4.1 describes the existing methods as Support Vector Machine as well as Pattern Deploying method. WBSVM is proposed method is better accuracy rate. While comparing the above two existing methods the proposed method gives high accuracy as well as its F-Measure value is also increased.

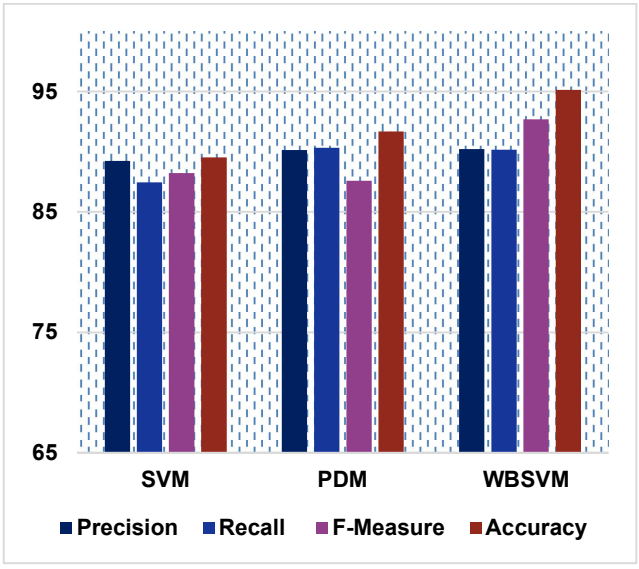

Fig.4.1 Web Based Support Vector Machine Pattern Taxonomy Classification Method

In Fig.4.1 explains the comparative chart on WBSVM classification method and existing methods. The proposed method WBSVM gives 95.12 % of accuracy and 92.67% value in F-measure. While comparing the existing methods the proposed method produce the high accuracy results.

VII. CONCLUSION

The various linguistic patterns performance and statistical scores is thoroughly analysed as well as assessed to develop a method that maximizes result quality. Our proposal is also reviewed across various well defined domains, consistently offering dependable taxonomies based on precision as well as recall. This paper initially focuses on developing an efficient classification algorithm for discovering patterns from extensive data collections. The pursuit of useful as well as intriguing patterns is emphasized. In the realm of text mining, pattern mining techniques are valuable for identifying different text patterns, including frequent itemsets, closed frequent itemset as well as co-occuring terms. The tools for web conent mining were necessary to analyse text as well as various document types, providing results to the search engine. The search engine is directed to deliver more effective search results by prioritizing their significance. This study examines the WBSVM method that evaluates various web content mining tools utilized to extract pertinent information from the associated web page.

REFERENCES

[1] Dr.S.Brindha, Dr.S.Sukumarn, “Relevance Pattern Discovery for Text Classification Using Taxonomy Methods” in International Journal for Science and Advance Research in Technology (IJSART)” Volume 4 Issue 11 –November 2018 ISSN [online]:2395-1052.

[2] Dr.S.Brindha, Dr.S.Sukumaran, “An Analysis on Big Data Interrogation Explore Problems and Tools”, International Journal of Novel Research and Development (IJNRD), ISSN:2456-4184, Volume 5, Issue 6,June 2020.

[3] Mustafa Ali Bamboat, GhulamSarfaraz Khan, NaadiyaMirbahar, SheebaMemon, “Web Content Mining Techniques for Structured Data: A Review” (SJHSE) Sindh Journal of Headways in Software Engineering. 2022;1(1).

[4]Richlin Selina Jebakumari, “A. Nancy Jasmine Goldena. A Survey on Web Content Mining Methods and Applications for Perfect Catch Responses”. International Research Journal of Engineering and Technology (IRJET). 2019;06(01): 407- 412. e-ISSN: 2395-0056 p-ISSN: 2395- 0072.

[5] Sharma PS, Yadav D, Thakur RN. “Web Page Ranking Using Web Mining Techniques: A Comprehensive Survey”. In M. P. Kumar Reddy (Ed.), Mobile Information Systems. 2022;2022:1–19. Hindawi Limited. Available:https://doi.org/10.1155/2022/751 9573.

[8] Kumar S, Kumar R. “A Study on Different Aspects of Web Mining and Research Issues”. In IOP Conference Series: Materials Science and Engineering. 2021;1022(1):012018. IOP Publishing. Available: https://doi.org/10.1088/1757- 899x/1022/1/012018.

[6] A. A. El-aziz, P. S. Pandian, S. N. Almuayqil, and A. S. Alruwaili, "A Framework for Clustering & Enhanced Approach for Frequent Patterns in Web Usage Mining A Framework for Clustering & Enhanced Approach for Frequent Patterns in Web Usage Mining", Assistant Professor , Department of Information Systems , College of Computer," no. June, 2020.

[7] S. Asadianfam, H. Kolivand, and S. Asadianfam, "A new approach for web usage mining using case based reasoning," SN Appl. Sci., vol. 2, no. 7, pp. 1–11, 2020, doi: 10.1007/s42452-020-3046-z.

[8] MinxiaoZhong, Yuqi Feng, Qing Li, Yanan Sun, "Precisely Predicting Neutronics Parameters of Nuclear\\xa0Reactor", *Advanced Intelligent Computing Technology and Applications*, vol.14863, pp.308, 2024.

[9] P. M. Bharti and T. J. Raval, "Improving Web Page Access Prediction using Web Usage Mining and Web Content Mining," Proceedings of the 3rd International Conference on Electronics and Communication and Aerospace Technology, ICECA 2019. pp. 1268–1273, 2019, doi: 10.1109/ICECA.2019.8821950.

[10] Pupale, R. (2019, February 11). Support Vector Machines (SVM)-An Overview. Medium. https://towardsdatascience.com/ https-medium-com-pupalerushikesh-svmf4b42800e989.

[11] R. Roy and G. Appa Rao, "Survey on pre-processing web log files in web usage mining," Int. J. Adv. Sci. Technol., vol. 29, no. 3 Special Issue, pp. 682–691, 2020.

[12] Samalikova, J., Kusters, R.J., Trienekens, J.J. and Weijters, A.J.M.M. (2014), "Process mining support for Capability Maturity Model Integration-based software process assessment, in principle and in practice", Journal of Software: Evolution and Process, Vol. 26 No. 7, pp. 714-728.

[13] Sarkar, P. (2019). Support Vector Machines in Machine Learning. https:// www.knowledgehut.com/blog/data-science/ support-vector-machines-in-machinelearning

[14] Yu, D., Xu, Z., & Wang, X. (2020). Bibliometric analysis of support vector machines research trend: A case study in China. International Journal of Machine Learning and Cybernetics, 11(3), 715–728. https:// doi.org/10.1007/s13042-019-01028-y

[15] M. Yeghaian, Z. Bodalal, T.M. TarecoBucho, I. Kurilova, C.U. Blank, E.F. Smit, M.S. van der Heijden, T.D.L. Nguyen-Kim, D. van den Broek, R.G.H. Beets-Tan, S. Trebeschi, "Integrated noninvasive diagnostics for prediction of survival in immunotherapy", *Immuno-Oncology and Technology*, vol.24, pp.100723, 2024.