Computer Systems Science & Engineering
DOI:10.32604/csse.2022.018300
Review

Tech Science Press

Ensemble Learning Models for Classification and Selection of Web Services: A Review

Muhammad Hasnain[1], Imran Ghani[2], Seung Ryul Jeong[3,*] and Aitizaz Ali[1]

[1]Monash University, Petaling Jaya, 46150, Malaysia
[2]Indiana University of Pennsylvania, Indiana, PA 15705, USA
[3]Kookmin University, Seoul, 02707, Korea
*Corresponding Author: Seung Ryul Jeong. Email: srjeong@kookmin.ac.kr
Received: 04 March 2021; Accepted: 30 April 2021

Abstract: This paper presents a review of the ensemble learning models proposed for web services classification, selection, and composition. Web service is an evolutionary research area, and ensemble learning has become a hot spot to assess web services' earlier mentioned aspects. The proposed research aims to review the state of art approaches performed on the interesting web services area. The literature on the research topic is examined using the preferred reporting items for systematic reviews and meta-analyses (PRISMA) as a research method. The study reveals an increasing trend of using ensemble learning in the chosen papers within the last ten years. Naïve Bayes (NB), Support Vector Machine' (SVM), and other classifiers were identified as widely explored in selected studies. Core analysis of web services classification suggests that web services' performance aspects can be investigated in future works. This paper also identified performance measuring metrics, including accuracy, precision, recall, and f-measure, widely used in the literature.

Keywords: Web services composition; quality improvement; class imbalance; machine learning

1 Introduction

Ensemble learning combines multiple classification techniques to achieve a high prediction accuracy than the single classification technique. In other words, an ensemble learning model is represented by the collection of different classification techniques and aggregated results to derive a best-fit model [1]. An ensemble learning model is either the same class of classification techniques or different classification techniques. The former type of ensemble is a homogenous ensemble learning model [2], and the latter type is known as a heterogeneous ensemble learning model [3]. For instance, a boosting classification technique similar to a 'random forest' (RF) technique can be used to build a more accurate and robust classification model that involves multiple iterations. As AdaBoostM1, one of the boosting classifiers undergoes more iterations, an ensemble learning model's predictive capacity is improved.

Ensemble learning models have been widely used in research areas, including bio-informatics [4], social media [5], software defect prediction, and classification [6]. Still, very little has been known about the applications of ensemble learning methods in web services' evolutionary research.

Web services selection regarding users' feedback helps determine users' trust. This trust of users is determined by evaluating the web services' quality attributes. The most well-known and widely explored quality attributes include response time, throughput, and reliability [7]. A user is always interested in selecting the best quality web service [8]. Quality attributes achieved at users' end need to be matched with the values mentioned in the 'service level agreement' (SLA) document [9]. The matching of users' achieved values and mentioned in the SLA document increases users' satisfaction. Different approaches have been proposed to evaluate web services classification and selection. To the best of our knowledge, this review is the first study that aims at reviewing the ensemble learning models and their applications in the area of web services.

The remainder of this study is organized as follows:

Section 2 describes the background; Section 3 describes the review method, Section 4 presents results and discussion; Section 5 concludes the study and presents a few implications.

2 Background

Web services are designed to enable the interoperable interaction between machines over the network. They are designed loosely for complex distributed software systems with a "service-oriented architecture" framework. The SOA framework is used to build services for end-users or integrate them with other services distributed over a network. With the help of SOA, there is a rapid growth in service-based systems [10]. Quality of Service (QoS) is the best criteria to investigate the performance of web services. Several factors affect the performance of web services. Cheng et al. [10] propose the approach to timely and accurately monitor the web services and their performance in a dynamic environment. Naïve based approach utilizes the sliding window and information gain theory to train the sample collected from web services.

The cost-sensitive and ensemble-based method (COSENS) framework [11] combines ensemble methods with cost-sensitive learning. The proposed framework has been evaluated on the risk factors of outsourced software projects. A low misclassification cost and a high accuracy indicated that the COSENS framework established the new rigorous evaluation standard. For the COSENS framework, the selection of optimal classifiers was undertaken by using a non-cost-sensitive situation. SVM1 remained the most optimal classifier based on the selection criteria with 74.9% accuracy compared to NB and 'decision tree' (DT) classification techniques.

Huda et al. [12] proposed an oversampling ensemble method to address the class imbalance issue in predicting software defects. The researchers in this study focused on reducing the consequences of high 'false negative' (FN) instances. Defect prediction based on the over accuracies has lesser impacts on reducing the error rate.

The 'average probability ensemble' (APE) [13] incorporates several classification techniques (RF, 'gradient boosting' (GB), 'stochastic gradient descent' (SGD), NB, and SVM). The proposed ensemble model for features selection showed a better performance in handling redundant features and irrelevant features. The performance of base classification techniques was done by averaging individual classifiers' performance that reduced the error rate. The proposed APF technique was evaluated on six publicly available datasets that outperformed the existing approaches based on the single SVM and RF classifiers.

Ensemble learning [14] is proposed from bagging and boosting classification techniques. Boosting techniques perform better for text classification than bagging techniques. Subsequently, the AdaBoost

classification technique is chosen for ensemble learning. Since the AdaBoost technique enhances a weak classifier's performance, the capability to process data is also increased. In the latter technique, different weak classifiers are trained on the dataset and then combine to achieve a robust classifier.

3 Review Methodology

To perform a rigorous review, preferred reporting items for systematic review and meta-analysis (PRISMA) guidelines [15] consisting of a 27-item checklist were used. The PRISMA flow chart is showing us the different phases of the review, such as identification, screening, eligibility, and included (see Fig. 1). PRISMA method helps researchers complete the reports for precise and rapid literature analysis. It may be helpful for critical analysis of the published review articles [15]. However, its checklist cannot be used to gauge the quality of a systematic review paper. The main difficulty of using the PRISMA method is the implicit expression of ways in the abstract and research method section of research studies.

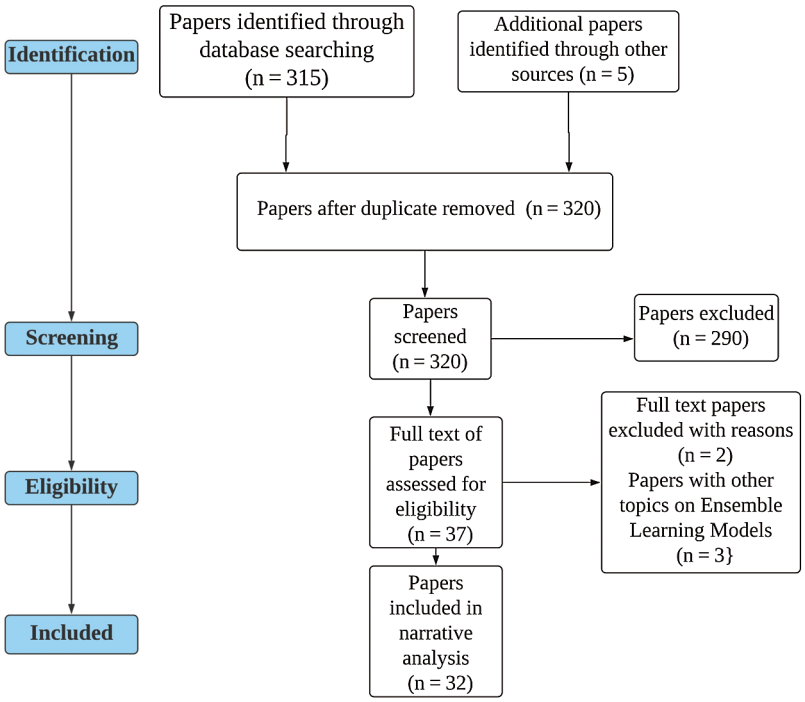

Figure 1: Review strategy

In the following, we present the search protocol adopted in each phase of the PRISMA guidelines to review ensembles learning models and their applications in web services, classification, and selection research areas.

Tab. 1 is the illustration of the search protocol (SP) adopted in this study. As listed in Tab. 1, we selected five popular digital libraries as data sources that covered machine learning and web services topics. Second, we constructed the search string based on the objectives of this study. Third, we applied eligibility criteria: (1) Search fields included title, abstract, and keywords; (2) Data coverage of studies ranged between January 1, 2010, and August 31, 2020; (3) Documents were Journal and conference papers; and (4) Language that considered studies published in English.

Table 1: Search protocol

SP	Data sources	ScienceDirect, IEEE Xplore, Taylor and Francis, ACM Digital Library, and Springer
	Search string	"Ensemble learning models" OR "Ensemble classification models" AND "Web services classification OR Web services selection."
	Search fields:	Title, abstract, keywords, and full contents
	Period	From January 1, 2010, to August 31, 2020
	Documents	Journal and conference papers
	Language	English

3.1 Research Questions

Based on the reviewed literature, we aim to investigate the state-of-the-art approaches on the ensemble learning models and web services. We proposed to define a few research questions as follows:

RQ1—What are state-of-the-art approaches that employed ensemble learning models in the context of web services classification and selection?

RQ2—What are the challenges to the ensemble learning models in the studied approaches?

RQ3—What are the performance metrics reported in the chosen studies?

3.2 Inclusion and Exclusion Criteria

Inclusion criteria

- We include papers that discuss "ensemble learning model" (ELM) along with web services themes.
- We include papers written only in the English language.

Exclusion Criteria

- We exclude papers with only abstracts.
- We exclude papers that discuss topics other than web services classification, selection, composition, and discovery.
- We exclude duplicate papers.

We report the four steps review strategy in Fig. 1. In the first step, we show the number of studies identified from search strings. In the next step, we perform the screening of papers and exclude those papers, which do not follow the studies' inclusion criteria. The subsequent step is focused on assessing the documents with the availability of full text. Only papers that follow the inclusion criteria are accepted.

3.3 Selected Studies

We include 38 papers in this review article. Of 32 papers, were published in Journal venue. The remaining papers were published across other venues. Tab. 2 presents the list of venues, quality or rank of venue, and the number of papers. We used two core ranks for Journal and Conference venues. The Journal venue was searched from SCImago Journal Rank, and the conference venue was searched from CORE Conference Portal.

Tab. 2 illustrates that most of the studies are published in Q1 quality Journals. This paper also includes a few quality conference papers.

Table 2: Name of venue, rank, and number of papers

Sr#	Name of venue	Rank	Papers
1	The Journal of Academic Librarianship	Q4	1
2	Journal of Software: Evolution and Process	Q3	1
3	Future Generation Computer Systems	Q1	3
4	PLoS One	Q1	1
5	Expert Systems with Applications	Q1	2
6	Applied Soft Computing	Q1	1
7	IEEE Transactions on Services Computing	Q1	1
8	International Conference on Developments in eSystems Engineering (DeSE)	C	2
9	Renewable Energy	Q1	1
10	Decision Support Systems	Q1	1
11	IEEE ACCESS	Q1	1
12	Information and Software Technology	Q1	1
13	International Parallel and Distributed Processing Symposium Workshops & Ph.D. Forum	A	1
14	Procedia Computer Science	–	1
15	International Journal of Data Science and Analytics	–	1
16	Business & Information Systems Engineering	Q1	1
17	NeuroImage	Q1	1
18	IEEE Journal of biomedical and health informatics	Q1	1
19	IEEE transactions on neural networks and learning systems	Q1	1
20	Proceedings of the 13th Innovations in Software Engineering Conference on Formerly known as India Software Engineering Conference	–	1
21	Mobile Networks and Applications	Q2	1
22	Journal of Discrete Mathematical Sciences and Cryptography	Q4	1
23	Computers & Electrical Engineering	Q1	1
24	Computers & Security	Q1	1
25	Journal of Systems and Software	Q1	1
26	International Journal of Information Technology	–	1
27	Social Network Analysis and Mining	Q1	1
28	Entropy	Q2	1

4 Results

This section presents answering the proposed RQs in this paper.

4.1 RQ1— What are State-of-the-Art Approaches That Employed Ensemble Learning Models in the Context of Web Services Classification and Selection?

To answer RQ1, we have investigated the chosen studies.

We find several studies that answer to RQ1. Singh et al. [16] proposed to use Bayesian Information Criteria (BIC) along with Akaike Information Criteria (AIC) as an evaluation metric for prediction of response time and throughput values. The latter mentioned approach used invocation time of web services with similar functionalities and then forecasted the best web services with convincing quality of services (QoS) values. Before this research, Raza et al. [17] proposed the weighted voting ensemble of eleven classifiers to select Amazon web services and Microsoft services providers. Earlier works did not present the criteria to know how users were satisfied with the web services. Users' feedback information of web services was further used to evaluate the performance of ensemble classification models. The logistic regression model showed better performance than the ensemble learning models.

Web services discovery [18] is another real challenge for researchers and software practitioners to develop service-oriented web applications. This is further compounded by the limited number of service management frameworks and intelligent categorization. To capture the functional semantics, researchers proposed a machine learning technique to overcome the earlier mentioned limitations. Ensemble learning classification categorizes the services into specific domains. AdaBoost, along with Random Forest and Bagging classifiers, were used. Combined techniques helped to capture the semantic relatedness between services. Accuracy results of ensemble learning models were better than the based estimators such as SVM and NB classifiers.

The proposed approach for the classification of web services uses significant vote-based ensemble classification models, including Naïve Bayes, SVM, and decision tree (J48) [19]. On the other hand, the proposed approach in Negi et al. [20] combined Naïve Bayes, Random Forest, and SVM classification techniques. The former technique used a decision tree classifier, and the latter used the Random Forest classifier. The rest of the two classifiers, such as Naïve Bayes and SVM, were the same for two ELM techniques. The ensemble learning technique resulted in a higher accuracy in comparison with the single technique. Higher accuracy results from ELM were also based on the improved preprocessing. An earlier approach [20] employed the Naïve Bayes, SVM, and REPTree classifiers in an ensemble learning fashion and showed accuracy improvement in the calibration of 'web services description language' (WSDL) documents. However, classification accuracy in Refs. [19,20] was better than in Laradji et al. [13]. This might be due to the maturity of ensemble learning techniques in the current era of machine learning.

Missing value and imbalanced data of attributes is another challenge in the bioinformatics domain. Negi et al. [20] proposed a service-oriented support decision system (SOSDS) based on the services-oriented architecture to overcome this issue. Web services represent the model for problem construction and related functionalities. Decision-making is performed by using lousy quality data and then assuming it to a binary classification problem. A committee of the SVMs is treated as an oracle that aims to correct training data and enables rule-based learning. Zięba [21] determined the prediction of unknown QoS values by building the K-nearest neighborhood of services based on their similarity results. So unknown values in the QoS matrix were generated using the known ones neighbors of the concerning services. A nonnegative latent factor (NLF) model performed better in predicting the missing QoS values with the high accuracy.

Anti-patterns are known as bad practicing solutions from the evolution or modification in web services with poor QoS values [22]. Earlier detection of anti-patterns in web services is performed by the ensemble learning method. The 'best training ensemble' (BTE) approach performed better than the majority voting ensemble' (MVE) technique.

Web services composition plays a significant role in satisfying the users' preferences [23]. Best composition patterns' mining with less time is still the central research area. The ensemble of the 'best first decision tree' (BFDT) and extreme boosting models is proposed in a recent research [24]. The

proposed approach is aimed to extract and mine the users' interesting patterns. Classification of web patterns into positive and negative correlations helps correct users' most interesting patterns.

A selective weighted voting ensemble model [25] has been proposed in recent research to improve the overall performance, such as variance, accuracy, and time consumption. Furthermore, a pruning technique is proposed using eigenvalues as an input weight matrix. The proposed technique's main objective was to reduce the computational cost and replace the single ELM network. The latter-mentioned approach applied to the random hidden parameters leads to the issue of low performance. The proposed approach is more effective as compared with the baseline ELM techniques. A homogenous ELMs based approach minimizes the number of operations and thus reducing the overall execution time.

Services quality and users' satisfaction are the best features of web services. Mohanty et al. [26] employed the hybrid ensemble learning technique (a combination of a neural network, TreeNet, SVM, and Decision Tree). To evaluate the proposed approach, experiments were performed on the quality of web services datasets. The proposed technique outperformed the existing approaches regarding average the accuracy. Web services feature selection yielded that throughput, response time, reliability, success-ability, and documentation were important. Among the above-mentioned classification models, if-then rule-based models such as CART and J48 were further suggested as expert systems for web services classification.

Sagayaraj et al. [27] propose selecting semantic web services (SWS) based on the ensemble machine learning approach. The selection of appropriate features is enabled by the maximum voting ensemble (MVE) approach. Top-ten web services were selected from different feature selection methods. The proposed ensemble machine learning approach outperformed the other methods. The MVE approach uses criteria to predict a class with the higher voting. Heredia et al. [28] proposed a method to recommend the web API based on APIs and mashups' trivial features. To achieve the objectives of web APIs' recommendation, the authors used the features ensemble and learning to the rank method. Text, nearest neighbor, API-specific, and Tag features were taken into account in this study. The proposed method was evaluated on web API datasets. Logistic regression, RF, RankNet models were used to analyze the data. However, the ensemble learning models remain inefficient to perform a deep understanding of several kinds of features.

The detection of fake reviews on the web is a complex task. It is essential for researchers to differentiate between truthful and untruthful reviews. Heredia et al. [28] used a combination of feature selection and ensemble techniques to detect spam reviews to address this issue. Bagging and boosting models were used as an ensemble technique to differentiate spam reviews from truthful reviews. Datasets were taken from three domains, including doctors, hotels, and restaurants. Still, we need to generalize the proposed technique on different datasets because current research employed text-based features.

Various social media and blog websites offer news access to users. The dissemination of fake news has become easier. The detection of fake news is a hot issue for researchers because existing supervised and unsupervised classification models show a poor accuracy towards the classification of news. Poor accuracy is the result of feature selection and their inefficient tuning and imbalanced datasets. Important feature selection in Hakak et al. [29] from fake news is performed with an ensemble learning approach. The decision tree, RF, and extra Tree Classifier show a higher accuracy than state-of-the-art machine learning approaches. Similar to Heredia et al. [28], this study needs to be evaluated on more datasets.

4.2 RQ2— What are the Challenges to the Ensemble Learning Models in the Studied Approaches?

Singh et al. [16] show a better prediction of the QoS metric values but does not show the efficiency than the recent popular approaches. The subjective and manual labeling of data proposed is time-consuming [30]. The subjective data labeling and its accuracy depend on the quality of labeling data. In Raza et al. [17], we

noticed that subjective labeling had time consumption and quality labeling issues. In other words, it is a multi-class single problem. It means that a single label for the review is assigned to a review, which has more than one aspect for the SaaS products. For instance, the slow response from services may be known as another aspect. Performance, scalability, and cost optimization best explain the services aspects. Therefore, we suggest researchers develop a precise method for labeling data that can reduce time in labeling large datasets.

The proposed approach [18] is efficient for the classification of web services with a static dataset. This approach shows limitations in managing the newly added services to a central repository known as 'universal description, discovery, and integration' (UDDI). Thus, automatic management may be activated to handle the changes occurring in web services. ELM proposed in Zięba [21] can help solve the class imbalance or missing values problem, but still, we need to address noisy labels.

Microblogging websites have remained a prime target of spammers. Twitter-spam has its impacts beyond social media. Many recent approaches used machine learning models to classify Twitter spam and have shown very promising results. However, many of these approaches have overlooked the class-imbalance issue for the real-world Twitter data. Liu et al. [31] propose an ensemble learning technique to accurately classify the redistributed dataset to overcome this issue. The majority voting scheme is combined with other classification models. The spam detection rate was improved with the imbalance class datasets.

Classification accuracies values reported in Negi et al. [20] are highly impressive in classification aspects. However, the application of the proposed ELM technique can be used to predict web services' performance. Based on the performance results, software practitioners may know their web services' future performance and correct if performance lowers than the promised QoS values. The proposed ensemble learning technique [32] can successfully predict false neighbors but still faces the challenges of analyzing the other QoS properties, and the selection of real-world services remains unresolved.

We can conclude that ensemble learning models still have the challenges of manual data handling, i.e., labeling, noisy labels, and time consumption. Therefore, future research can be undertaken with the support of deep learning models to investigate features deeply. Ensemble deep learning technique [33] is used to resolve the practical real life problems.

Trustworthiness is a key measure to decide the selection of wbe services. Multifaceted QoS attributes and their use for the selection of services is a problem. To overcome this issue, research in You et al. [34] introduced a selective ensemble learning approach. Particle swarm optimization (PSO) based technique measures the decimal weights while quantum discrete PSO calculates the binary weights. To evaluate the trust prediction approach based on the ensemble learning, public datasets were used. The proposed approach performed better than the current approaches of trust prediction. The key difference between this study and other research works is that data back-propagation is the latter-mentioned study's crucial method. Furthermore, You et al. [34] recommend the use of ensemble methods for classification with the higher accuracies.

4.3 RQ3— What are the Performance Metrics Reported in the Chosen Studies?

To answer RQ3, we have extensively studied the chosen papers. We have listed widely explored performance measuring metrics in the following Tab. 3.

Tab. 3 shows a list of ten performance measuring metrics studied in the chosen papers. The accuracy-based metrics such as accuracy, precision, recall, and f-measure were examined in studies [14,18,20,21,23] followed by MAE metric with three studies [22,23,32]. Sensitivity, Specificity, Cohen's Kappa, RMSE, and AUC metrics were used in each of two papers. Only two papers [22,23] have used the RMSE metric. Other than these metrics, Mathews' correlation coefficient and WPIA metrics were used in Raza et al. [17] and

Chakravarthy et al. [24] studies, respectively. This illustrates that accuracy-based metrics were mostly used in the literature.

Table 3: Performance measuring metric

Sr.#	Performance metrics	Reference
1	Sensitivity	[17,21]
2	Specificity	[17,21]
3	Cohen's Kappa	[17,20]
4	Matthews' Correlation Coefficient (MCC)	[17]
5	Recall Precision and F-Measure	[18,21,23,27]
6	Accuracy	[14,18,20,21,27]
7	Mean Absolute Error (MAE)	[22,23,32]
8	Root Mean Square Error (RMSE)	[22,23]
9	Area under the ROC Curve (AUC)	[23,28]
10	Web Pattern Identification Accuracy (WPIA)	[24]

Tab. 3 illustrates that Recall Precision, F-Measure, Accuracy, and Mean Absolute Error (MAE) are the most popular ones. In contrast, Matthews' Correlation Coefficient and Web Pattern Identification Accuracy (WPIA) are the least ones. F-Measure as an aggregate metric helps recognize the quality of problematic classes that better suit the problem of imbalanced data [35]. F-Measure too aggregates the simple sensitivity metric to determine the accuracy of classification of a minority class. In contrast, the precision metric indicates the probability of the correct prediction from F-Measure.

$$F1 - Score = 2 * \frac{Precision * Sensitivity}{Precision + Sensitivity} \tag{1}$$

MCC metric is derived from the confusion metrics measures such as true positive (TP), false positive (FP), true negative (TN) and false-negative (FN) as given in Eq. (2).

$$MCC = \frac{TP.TN - FP.FN}{\sqrt{(TP + FP)(FP + FN)(TN + FP)(TN + FN)}} \tag{2}$$

However, we cannot define the MCC metric if the value of any quantity of confusion matrix measure is zero [36]. Therefore, researchers choose the alternative metrics to compare the performance of their proposed ensemble learning models. Moreover, researchers in Ampomah et al. [37] called that F-Measure, Precision, Recall, and Accuracy metrics have a classical quality criterion used to quantify the performance of machine learning models.

WPIA metric is simply defined in the following Eq. (3)

$$WPIA = \frac{Identified\ users'\ interesting\ web\ patterns}{Total\ number\ of\ web\ patterns} * 100 \tag{3}$$

Eq. (3) is adopted from Chakravarthy et al. [24] where research is mainly focused on web services composition from interesting user patterns. This metric depends on other techniques to enhance its performance. Therefore, the later mentioned metric is specifically employed to determine the ratio between user web patterns and total web patterns.

5 Discussion

This section presents a discussion on the results obtained in this review paper. Fig. 2 below shows the publications trends in the chosen period.

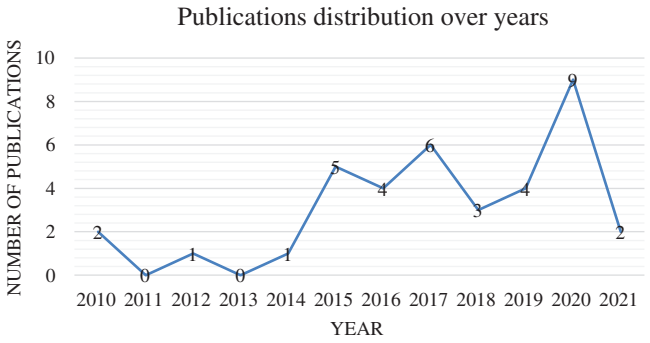

Figure 2: Research papers distribution over years (2010–2021)

Fig. 2 illustrates that the proposed ensemble learning methods and their applications increased dramatically. The history of the proposed ensemble learning methods is available for many years. Fig. 2 shows the evidence based distribution of research articles over the years. As shown in Fig. 2, from 2010 to now, ensemble learning methods have grown in selecting and classification web services. According to this section's findings, ensemble methods were used in two research articles in 2010, followed by nill in 2011. The results also illustrate that during 2010 and 2014, researchers have not shown much interest in this area of research. However, the number of publications increased to five papers in 2015. In addition, the number of research articles increased to 6 and 9 papers in 2017, and 2020 respectively. We have also observed that the ensemble learning methods have been widely used in different web applications, and it can be forecasted that this number may increase in the coming years.

Overall results of this review paper illustrated that the ensemble learning method showed prominent progress in web services, particularly web services classification and selection. With the help of ensemble methods, QoS prediction has enabled researchers to observe the time ahead performance of web services. Also, ensemble learning methods were mainly proposed from the well-known classification techniques such as SVM, Naïve Bayes, Random Forest, Decision Tree, AdaBoostM1, J48, etc., the classification and selection of web services. As given in Heredia et al. [28] that Logistic regression, RF, and RankNet models were used in a combination to analyze the data and rank the proposed method. Based on the feature selection and ensemble method, the study showed limitations in detecting the deep features of spam reviews.

This review paper identified some challenges regarding the applications of ensemble learning methods in web services research. Some of these challenges are common to other areas—for example, manual data labeling is a lengthy and time-consuming way that is still an open issue for researchers. As described in Raza et al. [17], the manual labeling of a large dataset is hard to manage the process. Similarly, ensemble methods [18] showed higher accuracies in classifying the static datasets of semantic web services. This method showed limitations in handling the dynamic QoS datasets of web services.

Fig. 3 indicates this paper's findings regarding the identification of performance measuring metrics. The majority of the proposed methods were evaluated using accuracy, Recall, Precision, and F-Measure metrics. Fig. 3 also provides evidence of the accuracy, and other metrics (Recall, Precision, and F-Measure) were used in 4 and 6 papers, respectively. The sensitivity metric was used in 3 articles. RMSE, MAE, Cohen's Kappa,

and specificity measures were used in two articles each. Each of WPIA, AUC, and MCC measuring metrics was employed in an article.

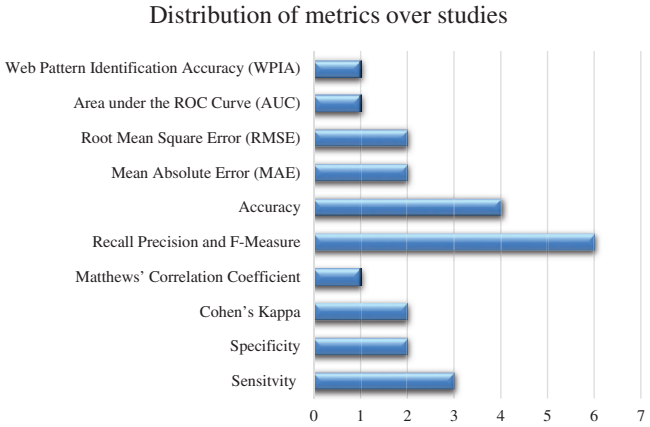

Figure 3: Distribution of performance metrics over publications

6 Conclusions

This paper presents an overview of the literature on the research topic of ensemble learning models and web services classification, selection, and composition. Advancement in machine learning technology introduces new challenges to web services' research area due to the imbalanced classes and missing values, decreasing classification models' accuracy. This paper shows an increasing trend in the ELM concerning web services quality improvement. In total, this paper identified ten performance measuring metrics. We grouped Precision, Recall, and F-measure metrics because the majority of studies report all three metrics.

The findings of this study have implications for researchers, web services developers, and managers. The results suggest that automated data labeling can save time and cost for larger web systems classification.

Funding Statement: This research was supported by the BK21 FOUR (Fostering Outstanding Universities for Research) funded by the Ministry of Education(MOE, Korea) and National Research Foundation of Korea(NRF).

Conflicts of Interest: The authors declare that they have no interest in reporting regarding the present study.

References

[1] K. W. Walker and Z. Jiang, "Application of adaptive boosting (AdaBoost) in demand-driven acquisition (DDA) prediction: A machine-learning approach," *Journal of Academic Librarianship*, vol. 45, no. 3, pp. 203–212, 2019.

[2] M. Hosni, A. Idri and A. Abran, "Evaluating filter fuzzy analogy homogenous ensembles for software development effort estimation," *Journal of Software: Evolution and Process*, vol. 31, no. 2, pp. e2117, 2019.

[3] M. Tang, X. Dai, J. Liu and J. Chen, "Towards a trust evaluation middleware for cloud service selection," *Future Generation Computer Systems*, vol. 74, pp. 302–312, 2017.

[4] A. Anaissi, M. Goyal, D. R. Catchpoole, A. Braytee and P. J. Kennedy, "Ensemble feature learning of genomic data using support vector machine," *PloS One*, vol. 11, no. 6, pp. 1–17, 2016.

[5] O. Araque, I. Corcuera-Platas, J. F. Sánchez-Rada and C. A. Iglesias, "Enhancing deep learning sentiment analysis with ensemble techniques in social applications," *Expert Systems with Applications*, vol. 77, pp. 236–246, 2017.

[6] Ö.F. Arar and K. Ayan, "Software defect prediction using cost-sensitive neural network," *Applied Soft Computing*, vol. 33, pp. 263–277, 2015.

[7] J. Liu, M. Tang, Z. Zheng, X. Liu and S. Lyu, "Location-aware and personalized collaborative filtering for web service recommendation," *IEEE Transactions on Services Computing*, vol. 9, no. 5, pp. 686–699, 2015.

[8] H. A. Kadhim and H. N. Nawaf, "Improve the accuracy of Dirichlet reputation system for web services," in *2018 11th Int. Conf. on Developments in eSystems Engineering (DeSE)*, Cambridge, United Kingdom, pp. 78–82, 2018.

[9] J. Heinermann and O. Kramer, "Machine learning ensembles for wind power prediction," *Renewable Energy*, vol. 89, pp. 671–679, 2016.

[10] H. Cheng, M. Zhong and J. Wang, "Diversified keyword search based web service composition," *Journal of Systems and Software*, vol. 163, pp. 110540, 2020.

[11] Y. Hu, B. Feng, X. Mo, X. Zhang, E. W. T. Nagi *et al.,* "Cost-sensitive and ensemble-based prediction model for outsourced software project risk prediction," *Decision Support Systems*, vol. 72, pp. 11–23, 2015.

[12] S. Huda, K. Liu, M. Abdelrazek, A. Ibrahim, S. Alyahya *et al.,* "An ensemble oversampling model for class imbalance problem in software defect prediction," *IEEE Access*, vol. 6, pp. 24184–24195, 2018.

[13] I. H. Laradji, M. Alshayeb and L. Ghouti, "Software defect prediction using ensemble learning on selected features," *Information and Software Technology*, vol. 58, pp. 388–402, 2015.

[14] L. Yuan-jie and C. Jian, "Web service classification based on automatic semantic annotation and ensemble learning," in *2012 IEEE 26th Int. Parallel and Distributed Processing Sym. Workshops & PhD Forum*, Shanghai, China, pp. 2274–2279, 2012.

[15] D. Moher, A. Liberati, J. Tetzlaff and D. G. Altman, "Preferred reporting items for systematic reviews and meta-analyses: The PRISMA statement," *International Journal of Surgery*, vol. 8, no. 5, pp. 336–341, 2010.

[16] V. P. Singh, M. K. Pandey, P. S. Singh and S. Karthikeyan, "An econometric time series forecasting framework for web services recommendation," *Procedia Computer Science*, vol. 167, pp. 1615–1625, 2020.

[17] M. Raza, F. K. Hussain, O. K. Hussain, M. Zhao and Z. ur Rehman , "A comparative analysis of machine learning models for quality pillar assessment of SaaS services by multi-class text classification of users' reviews," *Future Generation Computer Systems*, vol. 101, pp. 341–371, 2019.

[18] S. S. Kamath and V. Ananthanarayana, "Semantics-based web service classification using morphological analysis and ensemble learning techniques," *International Journal of Data Science and Analytics*, vol. 2, no. 1–2, pp. 61–74, 2016.

[19] U. Qamar, R. Niza, S. Bashir and F. H. Khan, "A majority vote based classifier ensemble for web service classification," *Business & Information Systems Engineering*, vol. 58, no. 4, pp. 249–259, 2016.

[20] N. Negi, P. Chaudhary and S. Chandra, "Web service classification based on non-functional parameters using vote based classifier," *Journal of Discrete Mathematical Sciences and Cryptography*, vol. 23, no. 1, pp. 45–55, 2020.

[21] M. Zięba, "Service-oriented medical system for supporting decisions with missing and imbalanced data," *IEEE Journal of Biomedical and Health Informatics*, vol. 18, no. 5, pp. 1533–1540, 2014.

[22] X. Luo, M. Zhou, Y. Xia, Q. Zhu, A. C. Ammari *et al.,* "Generating highly accurate predictions for missing QoS data via aggregating nonnegative latent factor models," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 27, no. 3, pp. 524–537, 2015.

[23] S. Tummalapalli, L. Kumar and N. B. Murthy, "Prediction of web service anti-patterns using aggregate software metrics and machine learning techniques," in *Proc. of the 13th Innovations in Software Engineering Conf. on Formerly Known as India Software Engineering Conf.*, Jabalpur, India, pp. 1–11, 2020.

[24] D. G. Chakravarthy and S. Kannimuthu, "Extreme gradient boost classification based interesting user patterns discovery for web service composition," *Mobile Networks and Applications*, vol. 24, no. 6, pp. 1883–1895, 2019.

[25] M. Abd-Shehab and N. Kahraman, "A weighted voting ensemble of efficient regularized extreme learning machine," *Computers & Electrical Engineering*, vol. 85, pp. 1–14, 2020.

[26] R. Mohanty, V. Ravi and M. R. Patra, "Web-services classification using intelligent techniques," *Expert Systems with Applications*, vol. 37, no. 7, pp. 5484–5490, 2010.

[27] S. Sagayaraj and M. Santhoshkumar, "Heterogeneous ensemble learning method for personalized semantic web service recommendation," *International Journal of Information Technology*, vol. 12, pp. 983–994, 2020.

[28] B. Heredia, T. M. Khoshgoftaar, J. D. Prusa and M. Crawford, "Improving detection of untrustworthy online reviews using ensemble learners combined with feature selection," *Social Network Analysis and Mining*, vol. 7, no. 1, pp. 1–18, 2017.

[29] S. Hakak, M. Alazab, S. Khan, T. R. Gadekallu, P. K. R. Maddikunta *et al.,* "An ensemble machine learning approach through effective feature extraction to classify fake news," *Future Generation Computer Systems*, vol. 117, pp. 47–58, 2021.

[30] M. S. Graham, I. Drobnjak and H. Zhang, "A supervised learning approach for diffusion MRI quality control with minimal training data," *NeuroImage*, vol. 178, pp. 668–676, 2018.

[31] S. Liu, Y. Wang, J. Zhang, C. Chen and Y. Xiang, "Addressing the class imbalance problem in Twitter spam detection using ensemble learning," *Computers & Security*, vol. 69, pp. 35–49, 2017.

[32] Y. Yin, Y. Xu, W. Xu, M. Gao, L. Yu *et al.,* "Collaborative service selection via ensemble learning in mixed mobile network environments," *Entropy*, vol. 19, no. 7, pp. 358, 2017.

[33] Z. Zhou, S. Tan, J. Zeng, H. Chen and S. Hong, "Ensemble deep learning features for real-world image steganalysis," *KSII Transactions on Internet and Information Systems (TIIS)*, vol. 14, no. 11, pp. 4557–4572, 2020.

[34] S. D. You, C. Liu and J. Lin, "Improvement of vocal detection accuracy using convolutional neural networks," *KSII Transactions on Internet and Information Systems*, vol. 15, no. 2, pp. 729–748, 2021.

[35] W. Wegier and P. Ksieniewicz, "Application of imbalanced data classification quality metrics as weighting methods of the ensemble data stream classification algorithms," *Entropy*, vol. 22, no. 8, pp. 1–17, 2020.

[36] S. Boughorbel, F. Jarray and M. El-Anbari, "Optimal classifier for imbalanced data using Matthews correlation coefficient metric," *PloS One*, vol. 12, no. 6, pp. 1–17, 2017.

[37] E. K. Ampomah, Z. Qin and G. Nyame, "Evaluation of tree-based ensemble machine learning models in predicting stock price direction of movement," *Information*, vol. 11, no. 6, pp. 1–22, 2020.