RESEARCH ARTICLE

DOI: 10.54392/irjmt25113

Asian Research Association

INTERNATIONAL RESEARCH JOURNAL OF
MULTIDISCIPLINARY TECHNOVATION

A Novel Hybrid Algorithm for Enhanced Re-Ranking Optimal Web Page Classification

J. Balaraju [a, *], S. Rahamat Basha [b], P. Ravinder Rao [a], P. Archana [c]

[a] Department of Computer Science Engineering, Anurag University, Hyderabad, India
[b] Department of Computer Science Engineering, Malla Reddy College of Engineering & Technology, Hyderabad, India
[c] Department of Artificial Intelligence, Anurag University, Hyderabad, India
* Corresponding Author Email: jb7443@gmail.com

DOI: https://doi.org/10.54392/irjmt25113
Received: 27-09-2024; Revised: 13-01-2025; Accepted: 19-01-2025; Published: 29-01-2025

Abstract: World Wide Web (WWW) is a platform that explores a wide range of information used for the development of web applications. Some examples of these applications include social network analysis, personalized item recommendations, and web page classification and ranking. Among these applications, search engines and web page ranking are particularly important as they consistently index and store billions of web pages on the internet. The main objective of this paper is to create an innovative framework for the classification and re-ranking of web pages using intelligent techniques. The framework is structured into two key phases: classification and re-ranking-based retrieval. In the initial classification phase, a series of pre-processing steps are implemented, including the elimination of HTML tags, punctuation, stop words, and the application of stemming. After pre-processing, a word-to-vector conversion is performed, followed by feature extraction utilizing Principal Component Analysis (PCA). This sequence of actions leads to optimal feature selection, which is vital for the precise classification of web pages. Given the multitude of features present in web pages that can compromise classification accuracy, this study employs a novel meta-heuristic algorithm, the Opposition Based-Tunicate Swarm Algorithm (O-TSA), to facilitate optimal feature selection. The refined features are subsequently processed through the Enhanced Convolutional-Recurrent Neural Network (E-CRNN), enhanced by O-TSA, resulting in the effective classification of diverse web page categories. In the second phase, the re-ranking process is executed using O-TSA, which establishes the objective function based on a similarity function (correlation) for URL matching, leading to optimal re-ranking of web.

Keywords: Correlation, Classification, ECRNN, OTSA, Principal Component Analysis

1. Introduction

As a result of the explosion of information currently accessible via the World Wide Web (WWW), numerous Web applications have been thoroughly investigated. Notable examples include Web page classification and ranking, social network analysis, and personalized item recommendation. Several of these uses, it is must to concentrate Web page ranking refers to how websites perform in search engine results. The billions of Web pages that search engines regularly trawl, store, and index [1]. They also offer a list of a great number of pages pertinent to a user's search, given a query from the user.

It is crucial to put the user-satisfying pages at the top of the list because not all of the pages in the list will pique the user's interest. This is the basic objective of the ranking algorithms used by search engines [2].

When a user submits a search query to the search engine, the relationship between the terms they have in mind is not shown. Consequently, the search engine ignores any semantic links between the keywords and displays websites where they are found using the specified keywords. More consideration needs to be paid to the selection process in order to produce recommendations and rankings that are more useful in relation to the needs of the customer [3]. These kinds of supplemental sources may consist of a variety of known and unknown knowledge dispersed throughout the environment because the inquiring web service will always have some degree of a related relationship. Because it is based on the essential contextual information, consumer usage behavior is a common way to be used in the selection process [4].

Numerous feature selection strategies and algorithms are offered for Web page classification among these techniques. However, only a small number of machine learning classifiers have been tested for classification [5]. This information development is consistently fast so getting helpful data from mass information is testing and difficult. Associations are building information distribution centres for holding their recorded data in deliberate way [6].

The undertaking of asset discovering is to extricate the crude data from the web reports. During the second undertaking, the extricated data is pre-processed by eliminating irregularity and clamour. Speculation intends to apply the example disclosure calculations to the pre-processed data to get the ideal examples [7]. Web mining is the general cycle of extricating, finding and breaking down data from the web information. The web mining process is segmented into three primary categories: web-content mining, structure mining, and usage mining [8].

Gate crashers make inconveniences to the users and keenly hack the mystery coding of them. In this way, the extortion discovery and control assume an essential function in the exploration territory and it is utilized by numerous analysts. Information mining strategies are viably applied in the research zone of extortion recognition and control. Misrepresentation discovery is a piece of the general cycle of misrepresentation control, which mechanizes and lessens the manual pieces of an assessment cycle. Visa conditional extortion discovery has got uncommon consideration for a large portion of the analysts. The research issues in information mining misuse data contained in printed records through different ways [9].

2. Deep Learning in Web Mining

An artificial neural network variation known as a "deep neural" can be identified by its depth of learning. There are multiple hidden layers; the reason for this is that the observable data is produced by the interaction of multiple layered causes. The layered factors in this network match the abstraction level [10]. Compared to artificial neural networks, deep neural networks contain extra properties. They are able to model complex non-linear relationships. Every layer of the deep neural network is trained using a unique set of features that are based on the results of the layer before it [11].

The WWW contains dynamic data, and the associations between the data have an unpredictable structure. Web mining is the term for the information mining of the WWW [12]. It contains heterogeneous and dynamic nature of data stockpiling zones. It manages three unique kinds of information like Web Substance or Content, Web Structure and Web Use data [13].

The web mining undertakings are ordered into the accompanying four stages Resource discovering, Information determination and preprocessing, generalization and, Analysis. The undertaking of asset discovering is to extricate the crude data from the web reports. During the second undertaking, the extricated data is preprocessed by eliminating irregularity and clamor. Speculation intends to apply the example disclosure calculations to the preprocessed data to get the ideal examples. During investigation, the examples are broken down for fulfilling the intriguing quality measure. Web mining is the general cycle of extricating, finding and breaking down data from the web information [14].

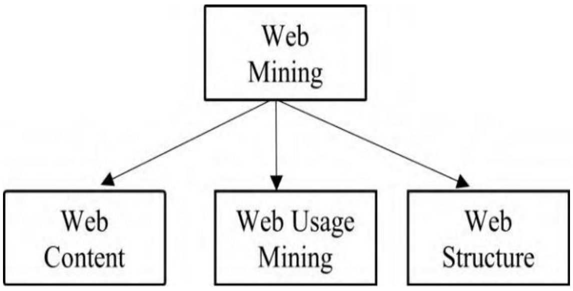

Figure 1. Web Mining Categories

This research aims to create an innovative smart webpage organization and rearrangement system. The proposed method enhances the effectiveness of sorting and reordering while reducing speed requirements and search times in web documents, differing from existing methods. This research intends to improve data quality by applying pre-treatment techniques to achieve optimal feature selection, thus augmenting classification accuracy via ECRNN, and ultimately assessing and comparing results with present approaches to ensure superior outcomes.

3. Problem Definition

Due to the difficulties and assortment of uses related with the issue of finding fascinating information from enormous genuine data sets, creators investigated their research in science, designing, data hypothesis, machine taking in and insights from information mining perspective. The information mining research covers the inescapable zones, for example, information stream, spatial, worldly, time arrangement, text, misrepresentation location and WWW. The accompanying research works demonstrated the difficulties and headways of information mining research progress by number of analysts. One of the difficult and significant research zones in information mining is to deal with high Text mining research is an interdisciplinary research field which removes valuable data from the printed reports. Web is a significant vehicle for data search and business exchanges. It has unstructured information design as markup dialects. Web mining assists with finding the information from the unstructured information store and coordinates the information in assortment of uses. The improvement of WWW creates huge volume of electronic data. The core elements of search engines are represented by web page rankings. The major goal of ranking is to produce a ranked list of Web pages that the user prefers best in response to a given user query. The three types of ranking algorithms are content-oriented, link-oriented, and hybrid strategies. Since these are the well-known and well-studied methodologies, the hybrid ranking algorithms make use of both the link and the content information. For use in dynamic web contexts, incremental C-Rank does not experience accuracy loss.

4. Literature Survey

Koo *et al* [15]. Have attempted to overcome C-Rank's restriction. It was suggested to use an incremental C-Rank, which was intended to update the C-Rank rankings of only a selective subset of Web pages rather than all of them with no loss of accuracy. The usefulness and efficiency of this suggested strategy were proven by experimental findings on a real-world dataset. They also provide a list of numerous pages relevant to a user's search, based on a query from the user. Because not every page in the list will grab the user's attention, it is important to place the pages that satisfy the user at the top. This is the primary goal of search engine ranking algorithms.

Chahal *et al* [16] are put forth a semantic web document ranking scheme that took into account both the conceptual instances that exist between the keywords as well as the keywords themselves. As a result, just the pertinent page would appear at the top of the list of results for the search. All pertinent relationships between the keywords were investigated to better understand the user's purpose, and their significance was determined by calculating the percentage of these relationships on each web page. Compared to the previously used techniques, this ranking technique produced better results.

Rong *et al* [17]. are recommended a method of private profiling to enhance the performance of ranking and suggestion. Since the composition process had a significant impact on service selection, personal knowledge from earlier service composition processes was gathered and shared via collaborative filtering, which initially identified a group of users with comparable interests. A web service re-ranking technique was subsequently used for customised suggestion. Experimental experiments were carried out and examined to show this research's promising potential.

Michal *et al* [18] proposes as a piece of a test, we have executed the entirety of the principal SEO procedures to a site, which offers traveller data about Bratislava in English. We have arranged a rundown of catchphrases with high volume of searches and low rivalry, whereupon we have played out the advancement steps. Techniques for search engine optimization, or SEO for short, ought to result in top spots in natural search results. The foundation of SEO still consists of certain optimization strategies that remain constant over time. However, new optimization strategies come and go as the Internet and web design change constantly. Therefore, we will examine the most crucial elements that can enhance a position in search results. It is crucial to stress that none of the methods can ensure it because search engines use complex algorithms to evaluate web pages' quality and determine where they appear in search results. The optimization's object, a specific website, is then introduced and examined.

Amran *et al* [19] compares mining algorithms, web usage mining, and numerous factors to take into account when looking for patterns. To meet future user expectations, quality information is required because enormous amounts of data are being added to repositories every second. In the future, user interest should be taken into account when establishing Web Services and discovering Web Services. This paper's primary goal is to determine whether search engine optimization raises a website's ranking in search results, which in turn increases traffic. Tests and results verification support this research question. The research findings are concluded and future directions are suggested in the last section of our paper.

Samedin Krrabaj *et al* [20], consider and describe Google SEO On-page and Off-page techniques. Our research focuses on the website www.studying-in-germany.org, and it uses both on-page and off-page SEO elements. Estimates are made using Google's SEERP measures and tools in order to consider the overall impact of various variables. At long last, we close the research and give future SEO bearing that will improve site positioning in web crawlers further by utilizing other SEO procedures. The Squirrel Search Algorithm (SSA), which takes into account the foraging behavior of flying squirrels in pursuit of food, is used to design the suggested SqSRank algorithm. The suggested ranking algorithm performs well in the web page re-ranking process thanks to the filtered web pages. By extracting the features linked to the web documents, the re-ranking measure is able to use the fitness measure to determine the ranking score. The optimal solution is acknowledged to be the fitness with the smallest retrieval distance.

Xi, Yunjia *et al* [21] are developed the headway approach of interest application system is finished by summing up measure and perceiving the capriciousness of the comparable strategies; and arranging the plan of the structure utilizing configuration design. Inquiry logs screen the information concerning interface among users and search motor. The transformation of the standard Page Rank check by utilizing 'weight 'of in-associated site pages. Instead of consistently isolating the weight of an in-connected website page, our technique appropriates it to all the out associated pages dependent on their ubiquity

5. Proposed System

To develop an intelligent framework for classifying and re-ranking web pages is the aim of this research work.

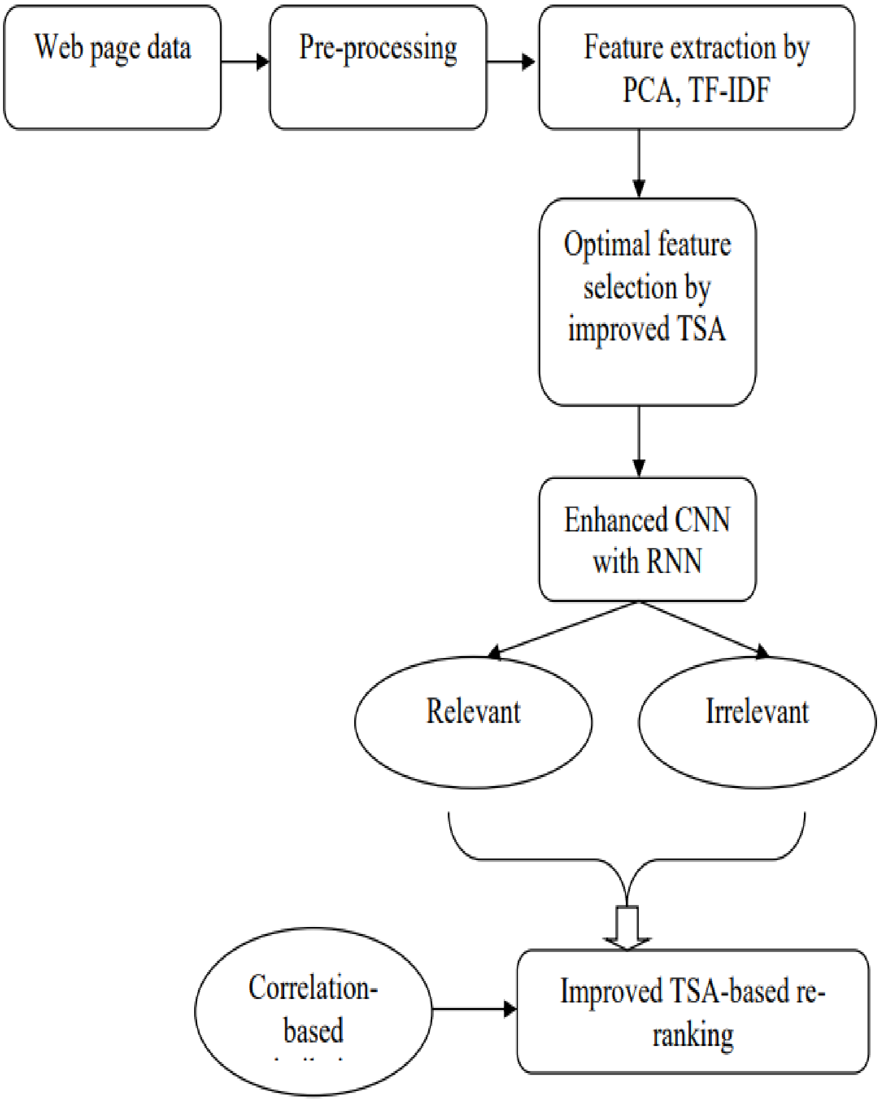
Web page data
Pre-processing
Feature extraction by PCA, TF-IDF
Optimal feature selection by improved TSA
Enhanced CNN with RNN
Relevant
Irrelevant
Correlation-based
Improved TSA-based re-ranking

Figure 2. Architecture of Proposed System

Layout of the system architecture is shown in Figure 2. The suggested model will consist of two basic phases: Classification & Re-ranking. Pre-processing, which includes actions like removing HTML tags, punctuation marks, common terms, stemming words, and stop words, will be carried out first in the categorization phase.

Upon the completion of per-processing, features will be extracted using Principal Component Analysis (PCA). This will result in the best feature selection, which is a crucial step in accurately classifying web pages. Web pages have a variety of properties, and adding more features lowers classification accuracy. The best feature selection in this case will be accomplished by using the new meta-heuristic technique known as the Tunicate Swarm Algorithm (TSA). When all is said and done, the chosen features subjected to the CNN with RNN in order to accurately classify web pages with improvements based on improved TSA. Classification of the web pages as relevant or irrelevant will be the result of this phase. The revised TSA, which will derive the Similarity-based objective function (correlation) while taking keyword matching, content matching, and URL matching into consideration, will be applied for the second phase's re-ranking, resulting in the best re-ranking of web sites.

Algorithm 1. Proposed OTSA

```
Initialize the swarm population
Declare RMIN = 1
Declare RMAX = 10
Declare swm = 0
While ( ) z < MAXi do
For NP i =1
Calculate the fitness of each search agents
If (b3 > 0.5)
Update the distance of food source based on the first constraint.
Update the swarm position based on best solutions of OTSA using first constraint.
else
Update the distance of the food source based on the second constraint.
Update the swarm position based on worst solutions of OTSA using the second constraint.
End if
End for
Update the final position of the tunicate using previous solution.
swm = 0
Update the parameters z = z +1
End while
Return best optimal solution FD
```

The suggested OTSA_CRNN utilizing data-driven machine learning methodologies across three distinct datasets with varying learning rates to assess the precision (98% of the proposed ideal web classification and reordering framework).

5.1. Accuracy of Proposed Model

At an 80 percent learning rate, the accuracy of the proposed OTSA-CRNN outperforms the, GWO-CRNN, PSO-CRNN, WOA-CRNN, & TSA-CRNN in the dataset no 1 by 6.18 percent, 2.06 percent, 5.67 percent, and 5.1 percent, respectively. At an 80 percent learning rate, the accuracy rate of the suggested OTSA-CRNN in dataset no 2 is 5.51 percent, 4.12 percent, 5.67 percent and 1.54 percent better than that of the PSO_CRNN, GWO_CRNN, WOA_CRNN, and TSA_CRNN, respectively. In the dataset no 3, the suggested OTSA_CRNN shows better performance than the NN, CNN, and RNN & CRNN in terms of accuracy at 5.15%, 10.3%, 7.21%, and 11.3% higher learning rates, respectively. As a result, the suggested optimum web categorization model, which combines the best feature selection and reranking, outperforms traditional models in terms of accuracy.

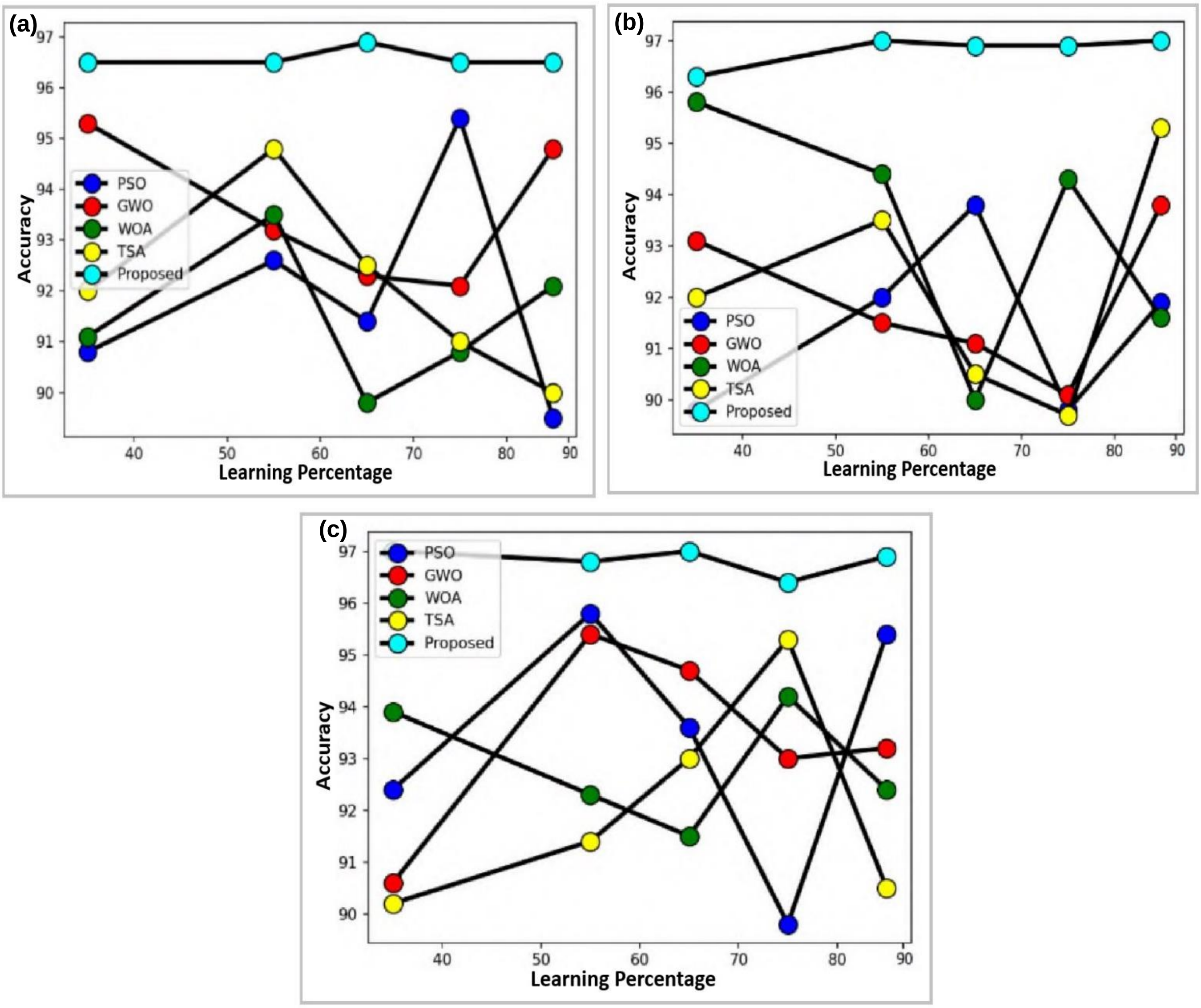

Figure 3. Accuracy analysis

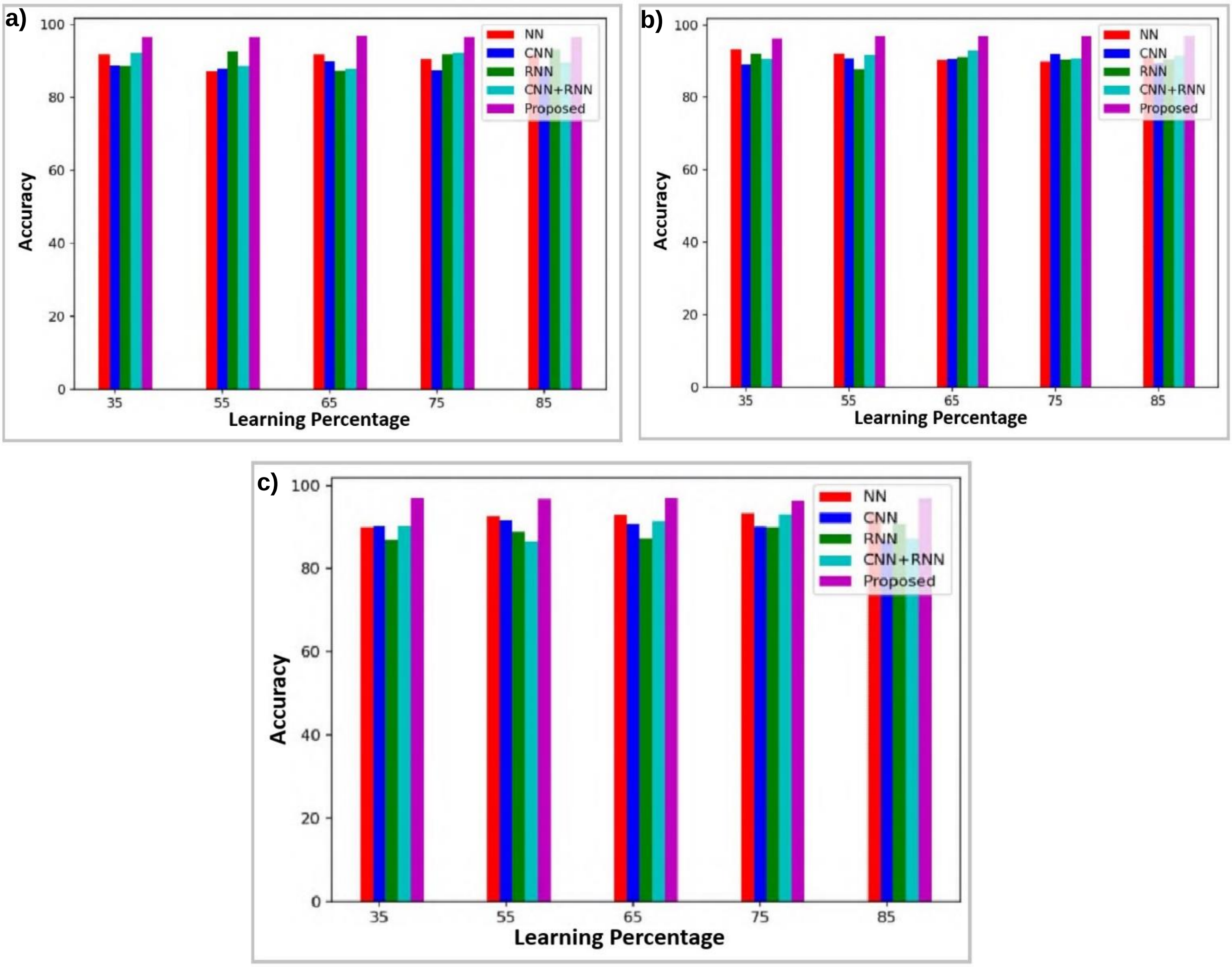

Figure 4. Accuracy analysis

6. Performance Analysis F1-Score

Using three datasets, which are depicted in Figure 3, the suggested method OTSA-CRNN is contrasted with help of ML methods. In the dataset no 1, the suggested OTSA_CRNN performs 3.03 percent, 4.04 percent, 6.06 percent, and 6.06 percent better than the other techniques, according to F1-score. When comparing the suggested OTSA_CRNN of dataset no 2 with the RNN, NN CNN, & CRNN, respectively, at the learning percentage of 65, the better F1-score is achieved as 3.03 percent, 4.04 percent, 6.0 percent, and 6.06 percent. The proposed OTSA-CRNN performs 3.03%, 4.04%, 6.06%, and 6.06% better than the, RNN, CNN, NN sand CRNN, respectively, in dataset 3 with an F1- score of 1. Thus, the suggested best web categorization model with the best feature selection has produced better results.

6.1 Precision Analysis for the Proposed Web -Classification Model

Using three datasets, Machine learning approaches are compared to the proposed OTSA_CRNN. in Figure 6. With a learning percentage of 65, the suggested OTSA- CRNN performs better than the CNN NN, RNN, , and CRNN in the dataset 3 by 1.01%, 2.02%, 2.02%, and 2.02%, respectively. As a result, the suggested best web categorization model with the inclusion of the best or optimal feature selection outperformed the traditional models in terms of precision.

6.2. Overall Performance Analysis

Three separate datasets are used to analyze the performance of the suggested web- page classification model utilizing the various heuristic-based algorithms as demonstrated in Table 1. The accuracy performance of the suggested OTSA_CRNN is 7.25 %, 1.76%, 4.5%, and 6.7% better than the PSO_CRNN, GWO_CRNN, WOA_CRNN, and TSA_CRNN, respectively, in the experimental analysis with dataset no 1.

When compared to PSO_CRNN, GWO_CRNN, WOA_CRNN, and TSA_CRNN, respectively, the suggested OTSA_CRNN is 0.33%, 0.45%, 0.6%, and 0.66% more advanced.

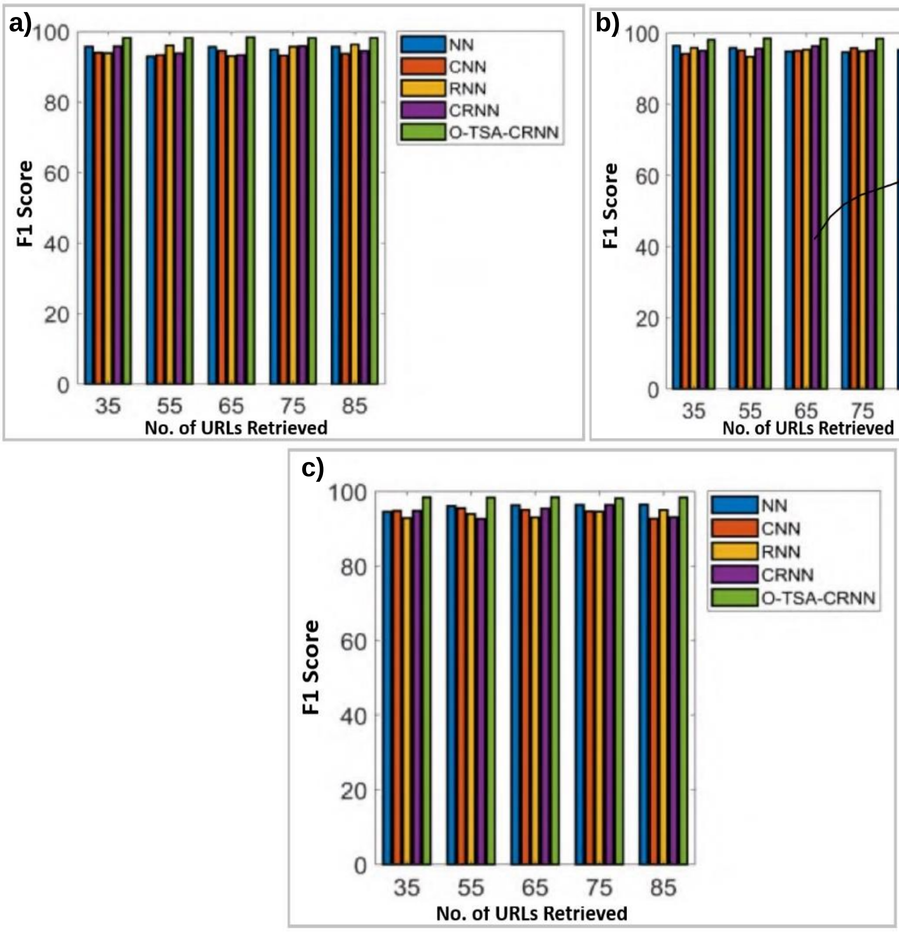

Figure 5. F1-score analysis

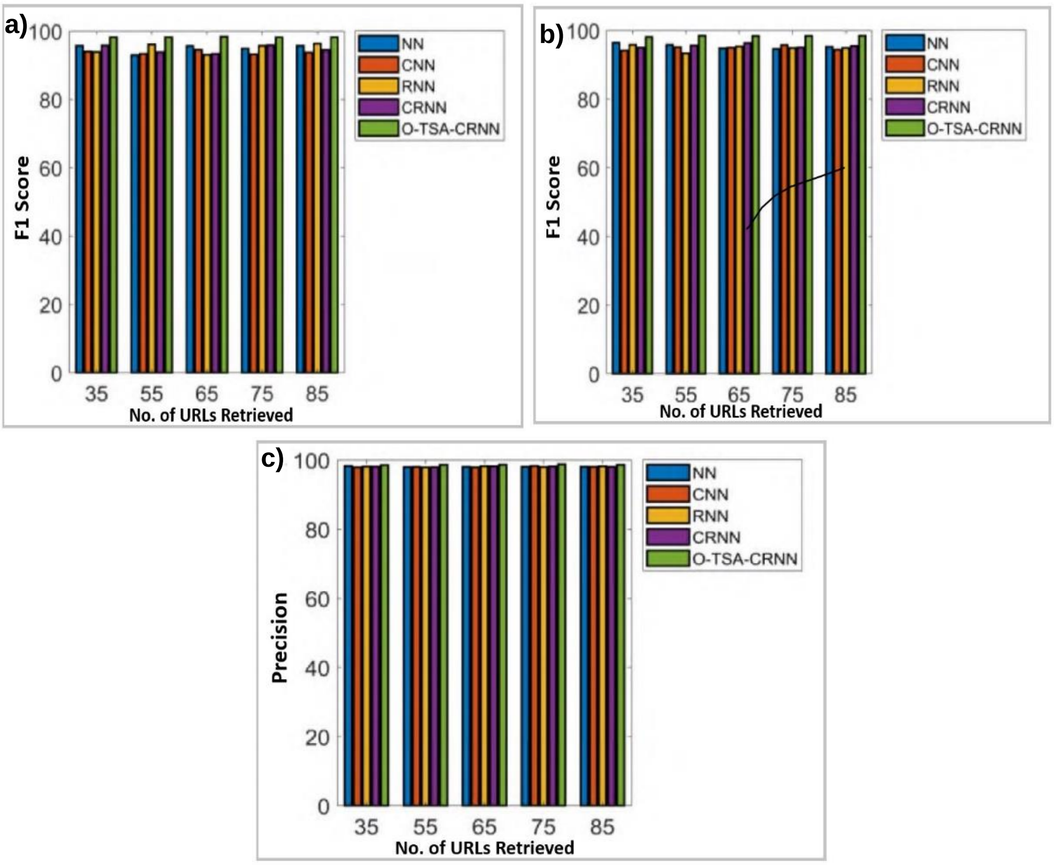

Figure 6. Precision analysis

Table 1. A comparison of three datasets using a suggested web classification

Dateset No 1					
Measures	**PSO-CRNN**	**GWO-CRNN**	**WOA-CRNN**	**TSA-CRNN**	**OTSA-CRNN**
Accuracy	0.895	0.948	0.921	0.9	0.965
Sensitivity	0.907	0.963	0.936	0.914	0.977
Specificity	0.512	0.514	0.534	0.512	0.500
Precision	0.983	0.982	0.980	0.98	0.986
FPR	0.512	0.485	0.508	0.505	0.500
FNR	0.09	0.036	0.063	0.085	0.022
NPV	0.500	0.514	0.500	0.512	0.500
FDR	0.016	0.017	0.019	0.02	0.013
F1-score	0.943	0.972	0.958	0.946	0.981
MCC	0.226	0.392	0.301	0.257	0.413
Dateset No 2					
Measures	**PSO-CRNN**	**GWO-CRNN**	**WOA-CRNN**	**TSA-CRNN**	**OTSA-CRNN**
Accuracy	0.919	0.938	0.916	0.953	0.97
Sensitivity	0.934	0.955	0.928	0.970	0.981
Specificity	0.513	0.457	0.516	0.454	0.52
Precision	0.980	0.979	0.983	0.981	0.987
FPR	0.486	0.542	0.483	0.545	0.481
FNR	0.065	0.044	0.071	0.029	0.018
NPV	0.513	0.457	0.516	0.454	0.52
FDR	0.019	0.020	0.016	0.018	0.012
F1-score	0.956	0.967	0.955	0.984	0.984
MCC	0.308	0.321	0.276	0.451	0.451
Dateset No 3					
Measures	**PSO-CRNN**	**GWO-CRNN**	**WOA-CRNN**	**TSA-CRNN**	**OTSA-CRNN**
Accuracy	0.954	0.932	0.924	0.905	0.969
Sensitivity	0.970	0.945	0.936	0.921	0.981
Specificity	0.468	0.516	0.531	0.486	0.48
Precision	0.982	0.983	0.983	0.979	0.986
FPR	0.531	0.483	0.468	0.513	0.521
FNR	0.029	0.054	0.063	0.078	0.018
NPV	0.468	0.516	0.531	0.486	0.48
FDR	0.017	0.016	0.016	0.020	0.013
F1-score	0.976	0.964	0.959	0.949	0.984
MCC	0.376	0.315	0.307	0.263	0.422

Table 2. A comparison of three datasets using a suggested web categorization model based on different classifiers

Dataset No 1					
Measures	**PSO-CRNN**	**GWO-CRNN**	**WOA-CRNN**	**TSA-CRNN**	**OTSA-CRNN**
Accuracy	0.19	0.884	0.931	0.897	0.965
Sensitivity	0.932	0.895	0.945	0.911	0.977
Specificity	0.5	0.5	0.529	0.513	0.5
Precision	0.98	0.983	0.982	0.979	0.986
FPR	0.522	0.512	0.470	0.486	0.510
FNR	0.067	0.104	0.054	0.088	0.022
NPV	0.512	0.523	0.529	0.513	0.500
FDR	0.017	0.016	0.017	0.020	0.013
F1-score	0.957	0.937	0.963	0.944	0.981
MCC	0.279	0.210	0.334	0.282	0.413
Dataset No 2					
Measures	**PSO-CRNN**	**GWO-CRNN**	**WOA-CRNN**	**TSA-CRNN**	**OTSA-CRNN**
Accuracy	0.91	0.895	0.904	0.914	0.97
Sensitivity	0.925	0.908	0.918	0.926	0.981
Specificity	0.485	0.534	0.468	0.516	0.52
Precision	0.980	0.982	0.981	0.983	0.987
FPR	0.514	0.523	0.531	0.483	0.480
FNR	0.074	0.091	0.081	0.073	0.018
NPV	0.485	0.512	0.468	0.516	0.521
FDR	0.019	0.017	0.018	0.016	0.012
F1-score	0.952	0.943	0.948	0.954	0.984
MCC	0.265	0.234	0.233	0.272	0.451
Dataset No 3					
Measures	**PSO-CRNN**	**GWO-CRNN**	**WOA-CRNN**	**TSA-CRNN**	**OTSA-CRNN**
Accuracy	0.933	0.866	0.907	0.873	0.969
Sensitivity	0.949	0.879	0.920	0.886	0.981
Specificity	0.485	0.484	0.515	0.484	0.481
Precision	0.980	0.9809	0.982	0.980	0.986
FPR	0.514	0.515	0.484	0.515	0.52
FNR	0.050	0.120	0.079	0.113	0.018
NPV	0.485	0.484	0.515	0.484	0.480
FDR	0.019	0.019	0.017	0.019	0.013
F1-score	0.964	0.926	0.950	0.931	0.984
MCC	0.321	0.191	0.266	0.199	0.422

The suggested OTSA_CRNN outperforms PSO_CRNN, GWO_CRNN, WOA_CRNN, and TSA_CRNN in the dataset 2 by 5.25 percent, 3.2 percent, 5.5%, and 1.75 percent, respectively, for the accuracy. By using FPR of the dataset 3, the suggested OTSA_CRNN is 2.11%, 9.85%, 6.9%, and 1.24% higher than RNN, NN, CNN, and CRNN.

As a result, the suggested optimum web-classification model with optimal-feature selection has outperformed traditional classifier models in terms of overall performance.

6.3. F1-Score Analysis on Page Retrieval

The result performance of the F1-score in the suggested optimal-web page retrieval model employing the proposed OTSA is evaluated across three datasets by varying the number of URLs retrieved, as shown in Figure 7.

The suggested OTSA performs 26%, 40%, 30%, and 20% better than the WOA, PSO, GWO, and TSA in the dataset 1 while obtaining 80 URL, correspondingly.

As a result, F1-score-based performance for page retrieval has surpassed that of traditional techniques.

6.4. Precision Analysis for the Proposed Web Page Retrieval

The suggested optimal web page retrieval model employing the proposed OTSA is evaluated based on precision with other heuristic-based approaches using three distinct datasets dependent on the number of URLs retrieved.

For fetching the 80 URL in dataset1, the suggested O-precision TSA's is 40%, 33.3%, 60%, and 36.6% better than that of the WOA, PSO, GWO, & TSA. On obtaining 60 URL in dataset-2, the suggested OTSA-CRNN method performs with precision that is 18.19%, 18.9%, 32.4%, and 8.18% better than the WOA, PSO, GWO, & TSA. By using dataset 3, the proposed OTSA outperforms WOA, PSO, GWO, and TSA in terms of 85 URL retrieval by 22.2%, 33.4%, 37.7%, and 11.11%, respectively.

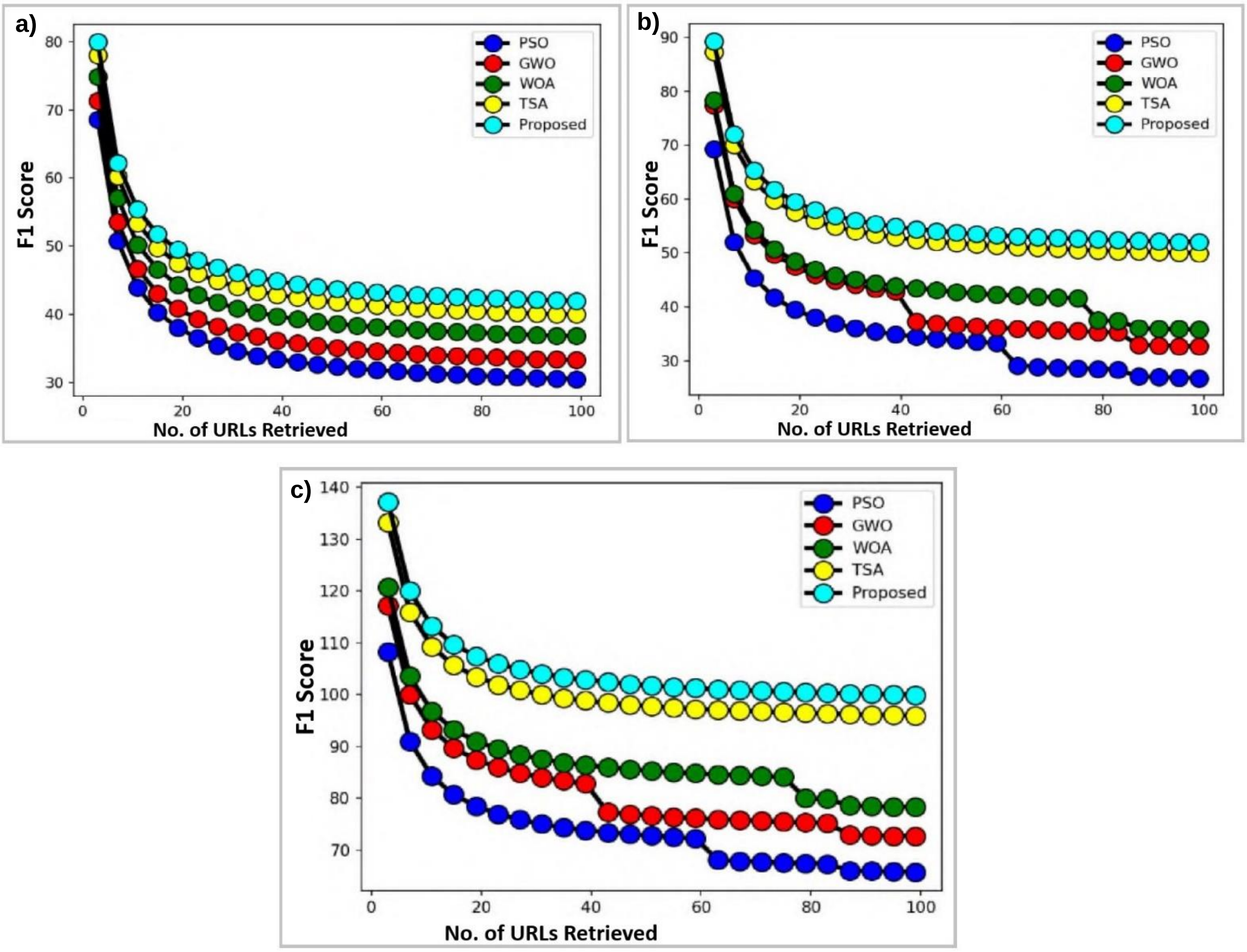

Figure 7. F1- Performance comparison

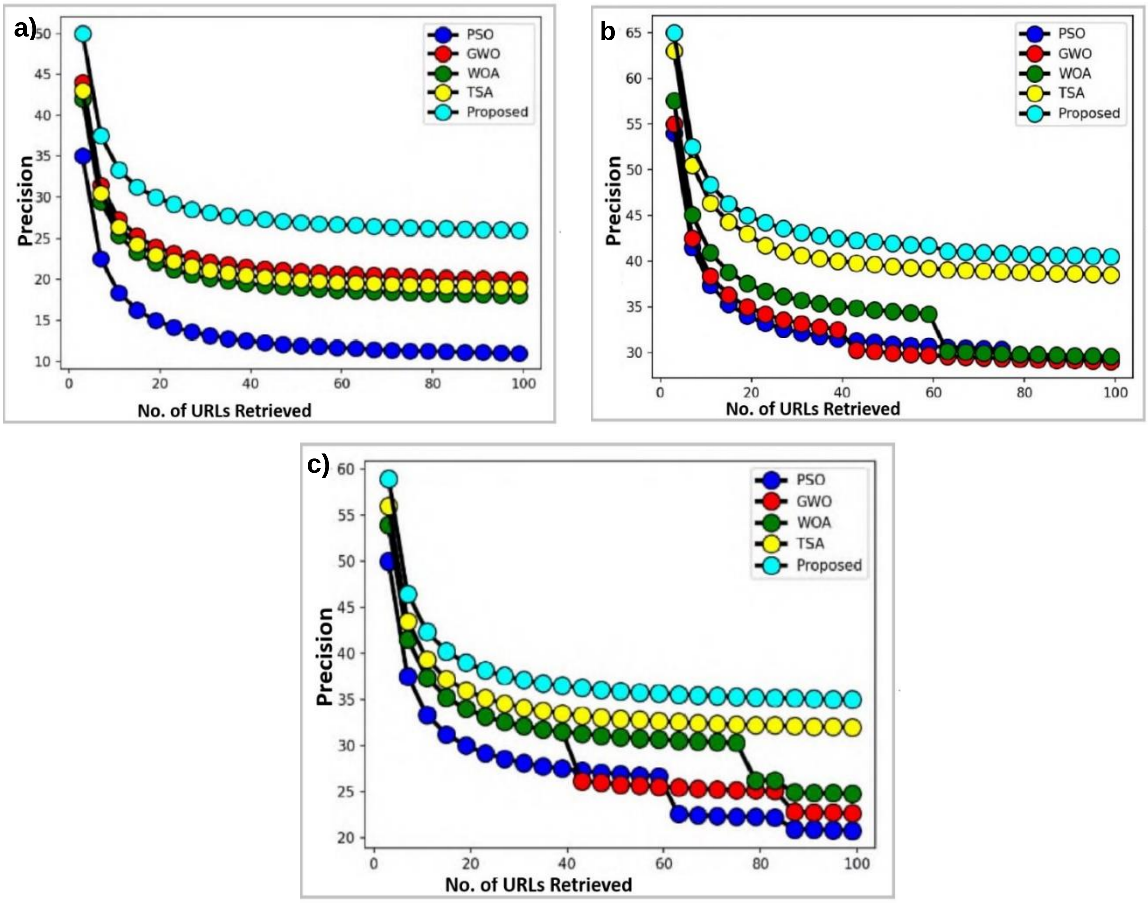

Figure 8. Precision Analysis

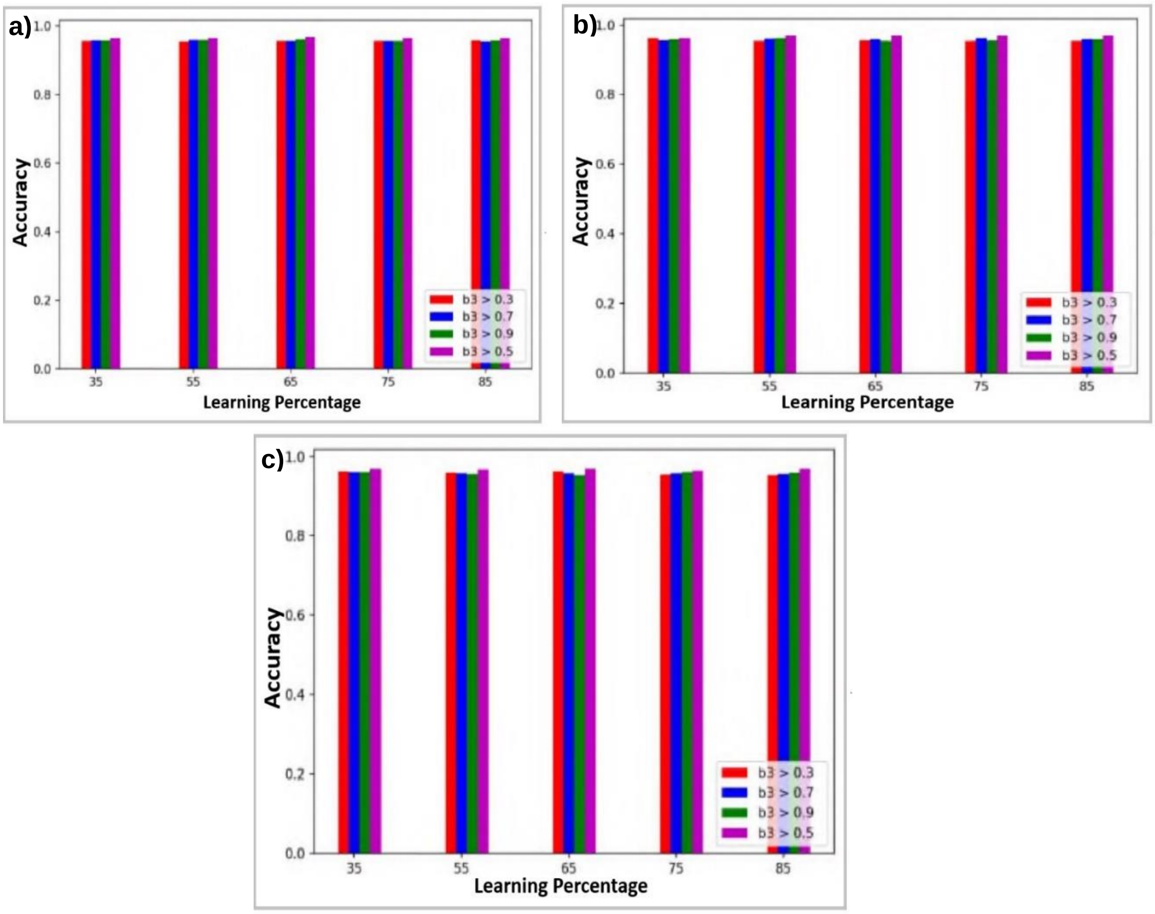

Figure 9. Proposed web page retrieval model

Table 3. Computation Time For The Three Datasets Used In The Proposed Web Categorization Model Based On Several Classifiers

Dataset 1	
Algorithm	Time (sec)
PSO- CRNN	152.9057 sec
GWO- CRNN	173.0744 sec
WOA- CRNN	149.2234 sec
TSA- CRNN	143.3077 sec
OTSA-CRNN	138.6211 sec
Dataset 2	
Algorithm	Time
PSO- CRNN	1664.384 sec
GWO- CRNN	1764.964 sec
WOA- CRNN	1616.125 sec
TSA- CRNN	1530.857 sec
OTSA-CRNN	1506.071 sec
Dataset 3	
PSO- CRNN	2446.811 sec
GWO- CRNN	2496.877 sec
WOA- CRNN	2397.45 sec
TSA- CRNN	2597.297 sec
OTSA-CRNN	2357.673 sec

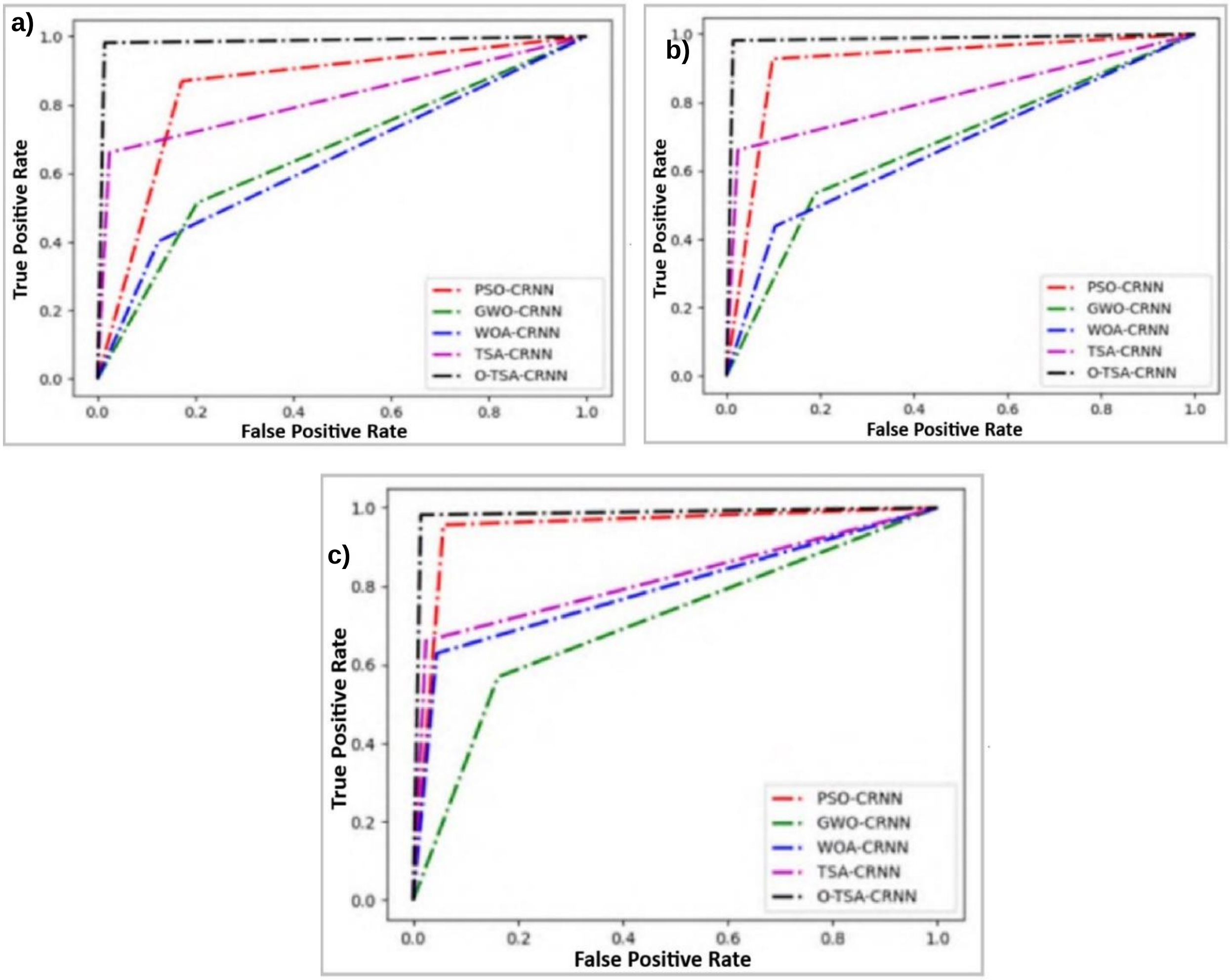

Figure 10. The suggested web page retrieval model's ROC curve for several techniques

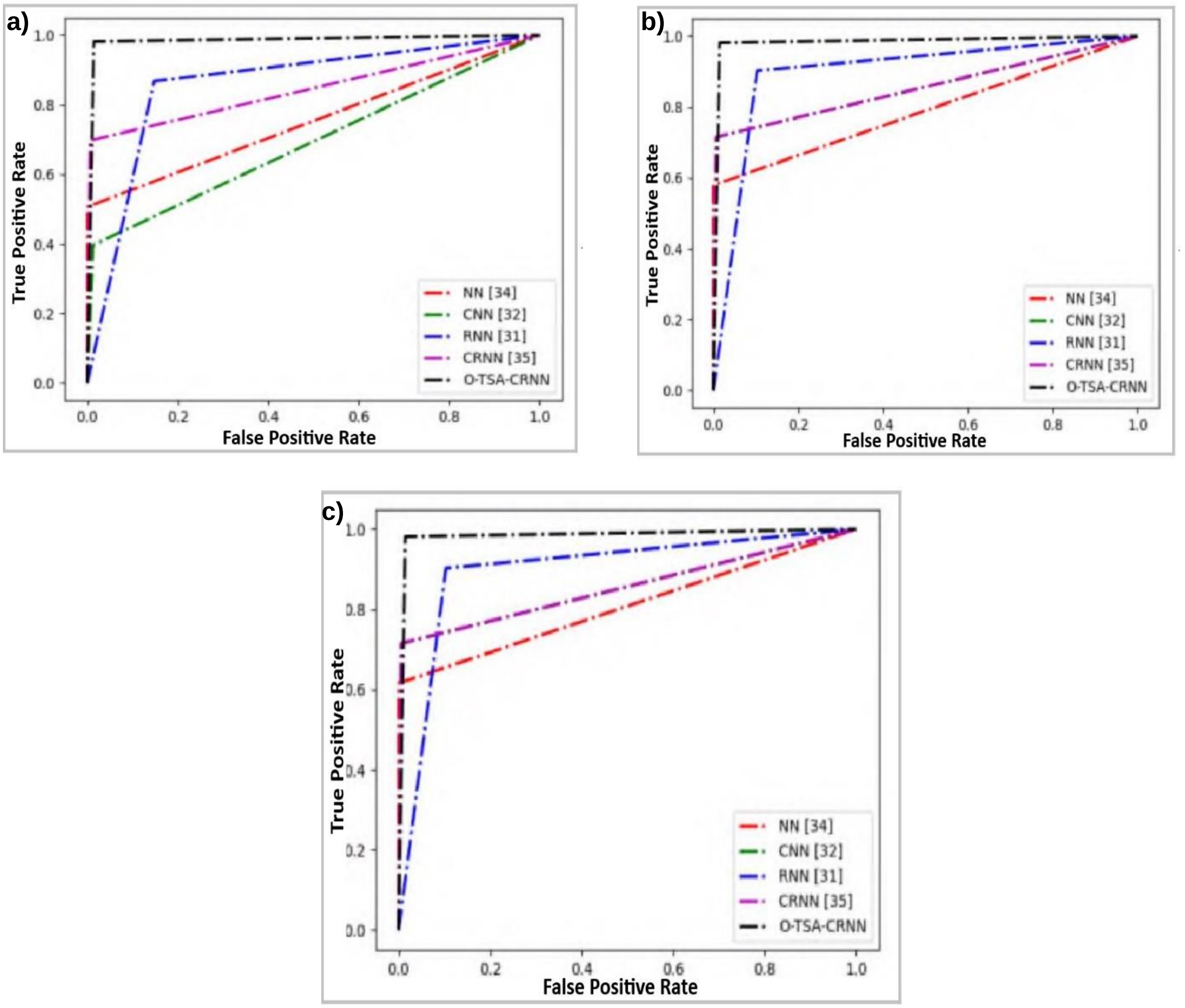

Figure 11. The suggested web page retrieval model's ROC curve for several classifiers

7. Web Page Retrieval Analysis Proposed Algorithm

Figure 8. Displays the analysis of the suggested algorithm. The results of the suggested OTSA with b3>0.5 is 12.5%, 6.25%, and 16.57% better than the b3>0.3, b3>0.7, and b3>0.9, respectively, at learning percentage55. As a result, the proposed OTSA with performs superior to the traditional methods.

7.1. Analysis on Computational Time

The Table 3 displays the suggested model's computational time. The suggested OTSA_CRNN requires less time as compared to PSO_CRNN, GWO_CRNN, WOA_CRNN, and OTSA_CRNN, respectively, by 9.34%, 19.90%, 7.10%, and 3.27%. Hence, compared to other current approaches, the suggested OTSA-CRNN requires the least amount of time to complete.

7.2. Roc Curve

The suggested web page retrieval model's Receiver Operating Characteristic (ROC) curve is shown in Figure 10 for a variety of approaches. With a false positive rate of 0.4 for dataset 1, the proposed OTSA_CRNN outperforms all other existing methods. The ROC curve for the suggested web page retrieval model for several classifiers is shown in Figure 11.

The suggested OTSA_CRNN outperforms the other well-known algorithms for dataset 2 with a false positive rate of 0.2. The proposed OTSA-CRNN performs better than the conventional methods in terms of ROC curve. Implemented OTSA_CRNN performs better than the established techniques.

8. Conclusion and Future Scope

To enhance the categorization and re-ordering of web pages, the most advanced web-based classification and re-ranking model, using the ECRNN classifier, has been executed in this research work. The input data from the web page underwent pre-processing and transformation into vector form as a result. Furthermore, features were derived through the PCA technique. The optimal selection of features for precise web page classification was achieved with the support of the proposed OTSA. Finally, the ECRNN was utilized on the selected features to accurately classify web pages by employing OTSA. The results were organized into categories based on URL. The re-ranking process, led by OTSA—which formulated the objective function utilizing a similarity function (correlation) concerning URL alignment with optimal weighted feature extraction—was conducted using the algorithm in a secondary phase, resulting in superior retrieval performance. The developed system faces certain challenges, such as the need to consider more features to identify optimal ones. In the future, any cutting-edge optimization algorithm may be integrated with deep learning techniques for enhanced classification and re-ranking.

References

[1] M. Gheisari, H. Hamidpour, Y. Liu, P. Saedi, A. Raza, A. Jalili, H. Rokhsati, R. Amin, Data mining techniques for web mining: a survey. In Artificial intelligence and applications, 1(1), (2023) 3-10. https://doi.org/10.47852/bonviewAIA2202290

[2] C. Choudhary, D. Mehrotra, A.K. Shrivastava, Enhancing the website usage using process mining. International Journal of Quality & Reliability Management, 41(9), (2024) 2311-2332. https://doi.org/10.1108/IJQRM-07-2022-0211

[3] B. Ravinder, S.K. Seeni, V. S. Prabhu, P. Asha, S. P. Maniraj, C. Srinivasan, (2024) Web Data Mining with Organized Contents Using Naive Bayes Algorithm. International Conference on Computer, Communication and Control (IC4), IEEE, India. https://doi.org/10.1109/IC457434.2024.10486403

[4] A. Breit, L. Waltersdorfer, F.J. Ekaputra, M. Sabou, A. Ekelhart, A. Iana, H. Paulheim, J. Portisch, A. Revenko, A.T. Teije, F. Van Harmelen, Combining machine learning and semantic web: A systematic mapping study. ACM Computing Surveys, 55(14s), (2023) 1-41. https://doi.org/10.1145/3586163

[5] J.K. Saini, D. Bansal, Computational techniques to counter terrorism: a systematic survey. Multimedia Tools and Applications, 83, (2024) 1189–1214. https://doi.org/10.1007/s11042-023-15545-0

[6] A. Dutt, M. Akmar Ismail, T. Herawan, I. Abaker Hashem, Partition-Based Clustering Algorithms Applied to Mixed Data for Educational Data Mining: A Survey From 1971 to 2024, IEEE Access, 12, (2024) 172923-172942. https://doi.org/10.1109/ACCESS.2024.3496929

[7] G. Papageorgiou, P. Economou, S. Bersimis, A method for optimizing text preprocessing and text classification using multiple cycles of learning with an application on shipbrokers emails. Journal of Applied Statistics, 51 (13), (2024) 2592–2626. https://doi.org/10.1080/02664763.2024.2307535

[8] P. Ristoski, (2023). Web Mining. Machine Learning for Data Science Handbook. Springer. https://doi.org/10.1007/978-3-031-24628-9_20

[9] M.S. Lin, R. Wen, (2024) A Web-based Text Mining System for Analyzing Customer Feedback of Returned Products. In Proceedings of the 2024 7th International Conference on Computers in Management and Business, 8-12. https://doi.org/10.1145/3647782.3647784

[10] S.H. Liao, R. Widowati, S.T. Liao, Two stages data mining analytics for food intentional and behavioral recommendations. Intelligent Data Analysis, (2024) 1-29. https://doi.org/10.3233/IDA-240664

[11] J.P. Bharadiya, A comparative study of business intelligence and artificial intelligence with big data analytics. American Journal of Artificial Intelligence, 7(1), (2023) 24. https://doi.org/10.11648/j.ajai.20230701.14

[12] A. Pradeep, (2023) Web Mining: Opportunities, Challenges, and Future Directions. International Conference on Intelligent Technologies (CONIT), IEEE, India. https://doi.org/10.1109/CONIT59222.2023.10205913

[13] S.P. Singh, M.A. Ansari, L. Kumar, (2023) Analysis of Website in Web Data Mining using Web Log Expert Tool. IEEE 12th International Conference on Communication Systems and Network Technologies (CSNT), Bhopal, India. https://doi.org/10.1109/CSNT57126.2023.10134696

[14] S. Ouf, Y. Helmy, M. Ashraf, Web Mining Techniques-A Framework to Enhance Customer Retention. International Journal of e-Collaboration (IJeC), 19(1), (2023) 1-30. https://doi.org/10.4018/IJeC.315790

[15] J. Koo, D.K. Chae, D.J. Kim, S.W. Kim, Incremental C-Rank: An effective and efficient ranking algorithm for dynamic Web environments. Knowledge-Based Systems, 176, (2019) 147-158. https://doi.org/10.1016/j.knosys.2019.03.034

[16] P. Chahal, M. Singh, S. Kumar, an Efficient Web Page Ranking for Semantic Web. Journal of the Institution of Engineers (India): Series B, 95, (2014) 15–21. https://doi.org/10.1007/s40031-014-0070-7

[17] W. Rong, B. Peng, Y. Ouyang, K. Liu, Z. Xiong,

Collaborative personal profiling for web service ranking and recommendation. Information Systems Frontiers, 17, (2015) 1265–1282. https://doi.org/10.1007/s10796-014-9495-4

[18] G. Michal, J. Zilincan, (2015) Improving Rank of a Website in Search Results-An Experimental Approach,10th International Conference on P2P, Parallel, Grid, Cloud and Internet Computing (3PGCIC), IEEE, Poland. https://doi.org/10.1109/3PGCIC.2015.145

[19] G.A. Amran, H.F. Aldheleai, H. Al-Sanabani, Understanding the Classification of Data Mining and Web Mining. International Journal of Applied Information Systems, 12(37), (2021) 36-39.

[20] S. Krrabaj, F. Baxhaku, D. Sadrijaj, (2017) Investigating search engine optimization techniques for effective ranking: A case study of an educational site. 2017 6th Mediterranean Conference on Embedded Computing (MECO), IEEE, Montenegro. https://doi.org/10.1109/MECO.2017.7977137

[21] Y. Xi, J. Lin, W. Liu, X. Dai, W. Zhang, R. Zhang, R. Tang, Y. Yu,. A bird's-eye view of reranking: from list level to page level. In Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining, (2023) 1075-1083. https://doi.org/10.1145/3539597.3570399

Authors Contribution Statement

J. Balaraju: Conceptualization, Implementation, Methodology, Experimental Result Finding, Writing – Original Paper Drafting and Editing, Visualization, and Paper Formatting. P. Archana: Data Pre-processing, Implementation, Experimental Result Finding, Accuracy Assessment, Fine Tuning and Optimization. P. Ravinder Rao: Conceptualization, Review of Existing Literatures, Writing -Paper Formatting, Editing, and Visualization. S. Rahamet Basha: Introduction, Exploratory Data Analysis, Writing –Editing, Paper Formatting, and Proof Reading. All the authors read and approved the final version of the manuscript.

Funding

The authors declare that no funds, grants or any other support were received during the preparation of this manuscript.

Competing Interests

The authors declare that there are no conflicts of interest regarding the publication of this manuscript.

Has this article screened for similarity?

Yes

Data Availability

The data supporting the findings of this study can be obtained from the corresponding author upon reasonable request.

About the License