The current issue and full text archive of this journal is available on Emerald Insight at:
https://www.emerald.com/insight/1742-7371.htm

Federate learning on Web browsing data with statically and machine learning technique

Ratnmala Nivrutti Bhimanpallewar
Department of Information Technology, Vishwakarma Institute of Information Technology, Pune, India

Sohail Imran Khan
Department of Business Administration, Lebanese French University, Erbil, Iraq

K. Bhavana Raj
Department of Management Studies, Institute of Public Enterprise, Hyderabad, India

Kamal Gulati and Narinder Bhasin
Amity School of Insurance, Banking and Actuarial Science, Amity University, Noida, India, and

Roop Raj
Education Department, Government of Haryana, Panchkula, India

Received 4 May 2022
Revised 7 July 2022
Accepted 21 July 2022

Abstract

Purpose – Federation analytics approaches are a present area of study that has already progressed beyond the analysis of metrics and counts. It is possible to acquire aggregated information about on-device data by training machine learning models using federated learning techniques without any of the raw data ever having to leave the devices in the issue. Web browser forensics research has been focused on individual Web browsers or architectural analysis of specific log files rather than on broad topics. This paper aims to propose major tools used for Web browser analysis.

Design/methodology/approach – Each kind of Web browser has its own unique set of features. This allows the user to choose their preferred browsers or to check out many browsers at once. If a forensic examiner has access to just one Web browser's log files, he/she makes it difficult to determine which sites a person has visited. The agent must thus be capable of analyzing all currently available Web browsers on a single workstation and doing an integrated study of various Web browsers.

Findings – Federated learning has emerged as a training paradigm in such settings. Web browser forensics research in general has focused on certain browsers or the computational modeling of specific log files. Internet users engage in a wide range of activities using an internet browser, such as searching for information and sending e-mails.

Originality/value – It is also essential that the investigator have access to user activity when conducting an inquiry. This data, which may be used to assess information retrieval activities, is very critical. In this paper, the authors purposed a major tool used for Web browser analysis. This study's proposed algorithm is capable of protecting data privacy effectively in real-world experiments.

Keywords Federation learning, Machine learning, GLM model, Web browsing, Data and privacy

Paper type Research paper

1742-7371
DOI 10.1108/IJPCC-05-2022-0184

Introduction

Federated learning is used to train a machine learning model, for example, on numerous local datasets housed in nodes with explicitly sharing data samples. The main approach is to train local networks on local sample data and to periodically exchange parameters (for example, the weights of a deep learning model) across these local nodes to produce a feature map recognized by all nodes. The fundamentals of browser-based forensics center on artifacts such as visited websites, malware URLs, time stamps, access counts, search history, cookies and downloaded activities. However, without the necessary precondition information, exploiting and identifying this information might be problematic. This article demonstrates how to do forensic analysis on common Web browsers like Chrome, Edge, Firefox, internet explorer and Dolphin on Android, as well as how a forensic investigator may gather forensic evidence from online browsers (Dhiman *et al.*, 2022). In this research, a consolidated image of all internet modes (public, private and portable) has now been constructed, together with powerful forensic qualities for the gathering of digital artifacts and tool comparison (Gulati *et al.*, 2021a). This article proposes a framework for gathering and analyzing evidence for Linux Web browser forensic. A methodology is provided for identifying questionable user behavior on the Internet (Gulati *et al.*, 2021b). Because an author's claim is a hypothesis, the author uses a hypothesis model to demonstrate their claim. Additionally, the author discusses the element of the customer's buying behavior for consumer items (Akanksha *et al.*, 2021). The author presents a relatively close semantics view analysis using a digital 3D model as well as a Web browser. The research is supported using two separate use cases: valuation of property investment and evaluation of green urban infrastructure (Virtanen *et al.*, 2021). After experimenting with chatbots in our personal lives, the author intends to use them in the office to aid us in choosing a new profession, resolving human resource concerns and even accepting coaching and mentorship (Wassan, 2021). This article describes the many reinforcement learning algorithms, their advantages and disadvantages, as well as the applications and difficulties that provide a path for future study (Akanksha *et al.*, 2021). The authors give a comparative examination of common machine learning-based classifiers in this research work. The author conducted experiments using tweet datasets of the COVID-19 pandemic. The author used seven classifiers based on machine learning (Wisetsri *et al.*, 2021).

This paper is organized into five sections. Section 1 includes introduction, the related study is described in Section 2, the methodology is presented in Section 3, Section 4 is analyzed results, and Section 5 includes a conclusion of the study and future work with limitations.

Related work

To solve crucial challenges like data privacy, security, access rights and access to diverse data, federated learning lets several actors work together to construct a single, strong machine learning technique without disclosing any of the underlying data. A wide range of businesses may benefit from it, including the military, telecommunications, internet-of-things (IoT) and pharmaceuticals, among others. The author is suggesting that a sequence of CNN models with varying depths provides diverse semantic features of the picture. According to test findings, IMCEC is an excellent tool for detecting malware (Akanksha *et al.*, 2021). Using CNN-based deep learning architecture, the authors present IMCFN, a new classifier for detecting malware variants (Gulati *et al.*, 2021a). The author presented the MLP model for fraud detection in secure e-banking e-commerce transactions for test websites (Dovhan *et al.*, 2021). In this article, the author presented the CNN model for plant and flowers detection (Billewar *et al.*, 2021). This study used a novel strategy that uses sentimental features centered on the item's attributes (Sanil *et al.*, 2021). The author reviewed and analyzed the literature, paying special attention to characteristics of wireless connections for energy conservation and data aggregation (Gulati *et al.*, 2021a). The main aim of this study is to scrutinize the conclusions of many studies and the advancement of e-participation in wealthy and poor India (Gulati and Telu, 2016). Image-guided surgery relies heavily on medical imaging, and in this article, the authors describe the most prevalent techniques for acquiring medical images (Alam *et al.*, 2018). The author has conducted an inquiry into energy utilization and stage-free administration in this study (PaaS). PaaS administrations are often used to supply phase administrations for application improvement (Bansal *et al.*, 2018). This work established an effective automated diagnostic system for maize plants. Data pre-processing, extraction of features, classification and segmentation are the four steps of the suggested technique (Akanksha *et al.*, 2021). The purpose of this research article is to address the use of intelligent machines (AI) to stock market modeling, demand planning and market segmentation challenges, with a particular emphasis on CNN models (CNN) and fuzzy logic. The first two issues were solved using backpropagation techniques, while the third was solved using self-organizing maps (SOM) (Joseph *et al.*, 2021). Author investigate the degree to which common Web browsers can withstand such assaults in this research by examining the pattern of their network activity while requesting webpages (Zhioua, 2015). The author evaluated over 1,000 browser security setting choices across three major browser and discovered that just 13 configurations had both semantic and syntactic similarity, whereas four configuration options shared only semantic similarity. Positive, negative and neutral effects of feedback were all included in the author's Recommendation Model (Sriram *et al.*, 2021). Author purposed a statically mode for sleepy student during the lecture also describe the sleeping impact on attention, working memory, mood, etc. (Wassan *et al.*, 2022). A Web browser may be used by a suspect to gather information, conceal unlawful activities and try out new criminal techniques. In digital forensic investigations, searching for evidence left behind by online surfing behavior is a common component. When a suspect uses a Web browser, he or she leaves a trail of evidence behind. The investigator may use this evidence to get insight into the suspect's personal information while doing so. You may use this evidence to examine online sites frequented by the suspect and their access times and frequency as well as search engine terms they use after retrieving data such as caches, history cookies and download lists from their machine. The main contribution of this work is to introduced a new tool for analyzing the Web browser data with different groups and views point. Our proposed algorithm is capable of protecting data privacy effectively in real-world experiments.

Methodology

Each kind of Web browser has its own unique set of features. This allows the user to choose their preferred browsers or to check out many browsers at once. If a forensic examiner has access to just one Web browser's log files, it makes it difficult to determine which sites a person has visited. The agent must thus be capable of analyzing all currently available Web browsers on a single workstation and doing an integrated study of various Web browsers. In this paper, we analyzed the Web browsing data with 9,000 samples, 4 special attributes, 12 regular attributes. the main aim of this paper is to create a new method for analyzing the Web browsing data and investigate the browsing history with different times and groups. We used the GLM model to predict better performance. In this paper, we used four steps. Figure 1 display the methodology process.

Step 1. We retrieve the Web data, set which attribute we want to predict (high/low value), and, finally, we remove those attributes which are very correlated and, therefore, do not provide

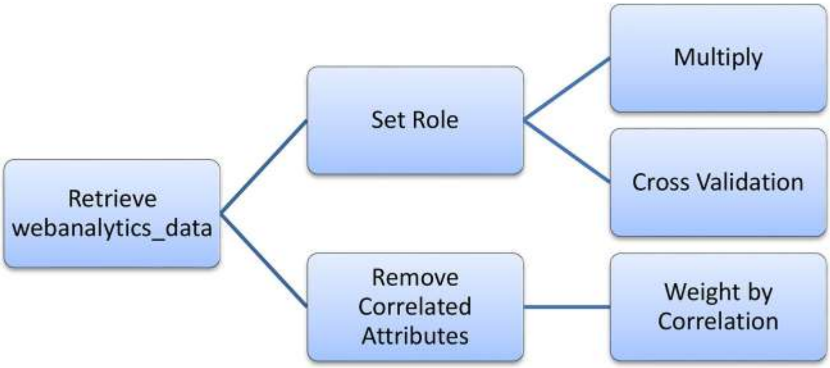

Figure 1.
Presented the methodology process

additional information. Then, in Step 2, data is copied to analyze it in two ways. In Step 3, generalized linear model is used to train a model and validate the prediction. The data is previously balanced to help the model to detect the less frequent “high value” cases. In Step 4, we compute the correlation weights to detect the most relevant attributes. Finally, we obtained the result. In Figure 2, we presented the true and false counts analyzed label information.

Analyzed results

In Figure 3, we presented the true label scenario with different groups which visited the websites at different times.

In Figure 4, we presented the false label scenario with different groups which visited the websites at different times.

In Figure 5, we presented the confidence true and false values with different groups which visited the websites at different times.

In Figure 6, we presented the count prediction high value with true count confidence and false count confidence label.

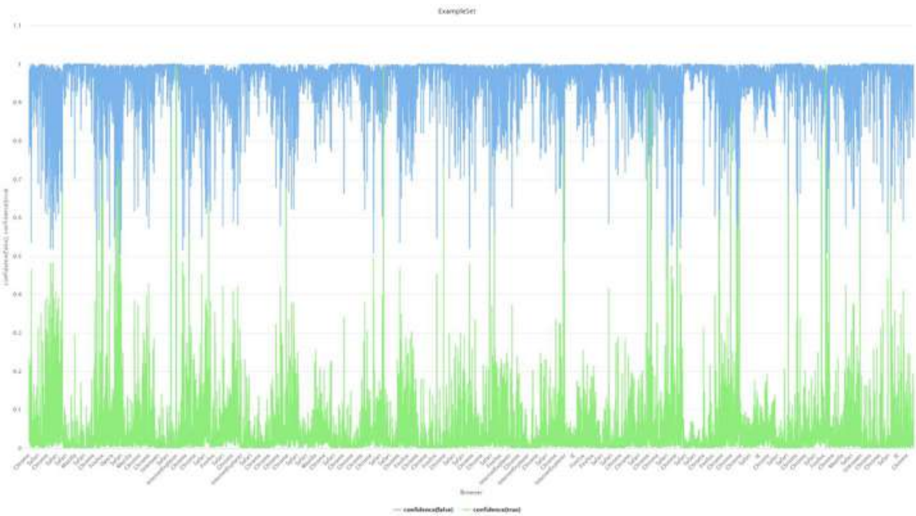

Figure 2.
Presented the true confidence and false confidence with 16 attributes

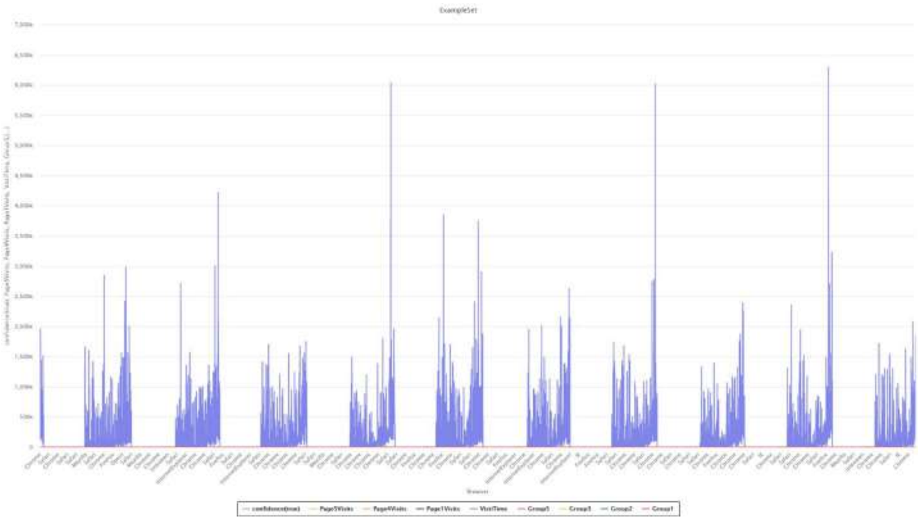

Figure 3.
Presented the true label

Table 1 represented the 12 attributes weighted results, where Groups 1–3 and Group 5 visited with different periods. Also, we presented the graphical view of weighted attributes in Figure 7.

Table 2 presented the statistical model with a GLM. There is a total of 16 attributes. The total attribute distributes with different sections, with 4 special attributes and 12 regular attributes. The co-efficient and stander deviation of all attributes are represented in Table 2.

In this, we presented the attributes with 16 features with different groups which visited the websites at different times.

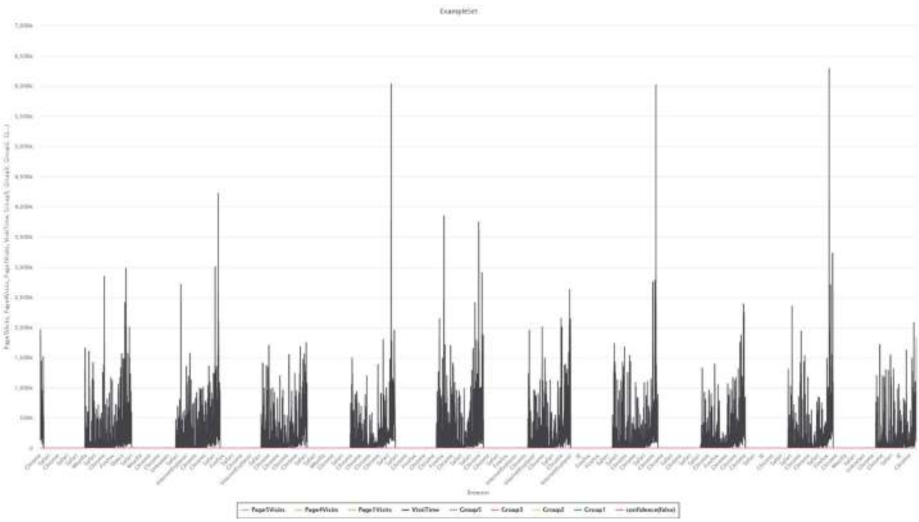

Figure 4.
Presented the false label scenario

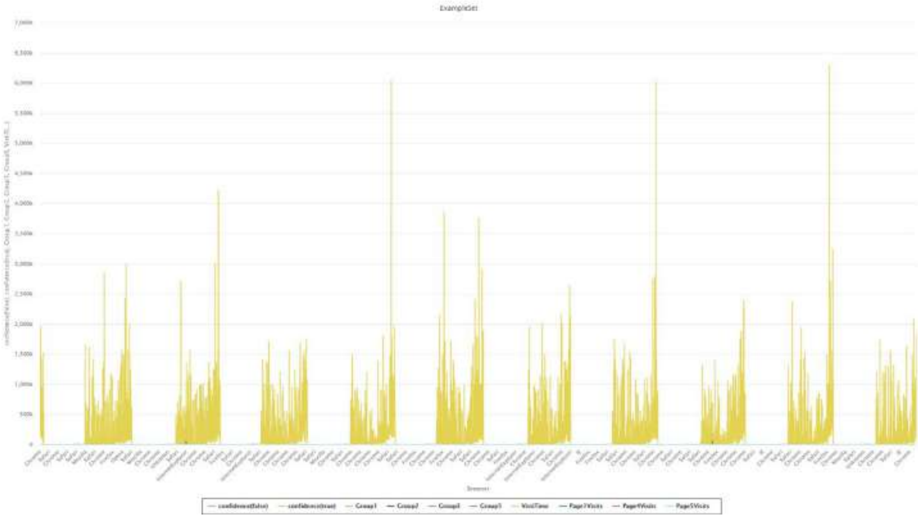

Figure 5.
Display the confidence true and false

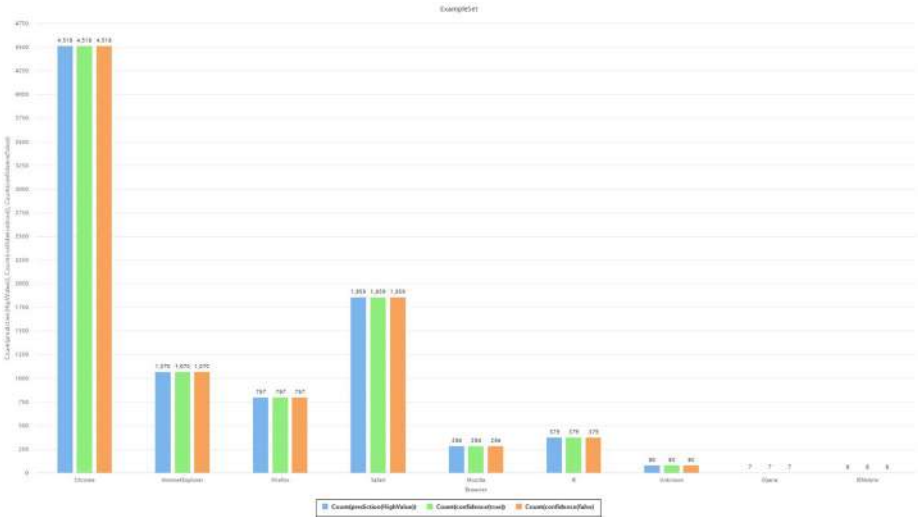

Figure 6.
Display the count prediction high value

Figure 8 heat map shows the static result of the high value predict by groups with different browser histories, like IE Mobile, Opera, Unicom, IE, Mozilla, Safari, Firefox internet and chrome.

Figure 9 heat map shows the static result of the high value predicted by visiting time with different browser histories, like IE Mobile, Opera, Unicom, IE, Mozilla, Safari, Firefox internet and chrome.

Federated learning with machine leaning technique

Federation machine-learning approach in which an algorithm is trained over a network of decentralized network edge or servers while maintaining local data samples without

Weight attributes result	
Visit time	0.067369
Reference	0.00907
Period	0.001238
Page5Visits	0.01368
Page4Visits	0.004299
Page1Visits	4.62E-04
Label	0.025893
Group5	0.006698
Group3	0.00847
Group2	0.026388
Group1	0.001779
Browser	0.001979

Table 1. Displays the weighted results

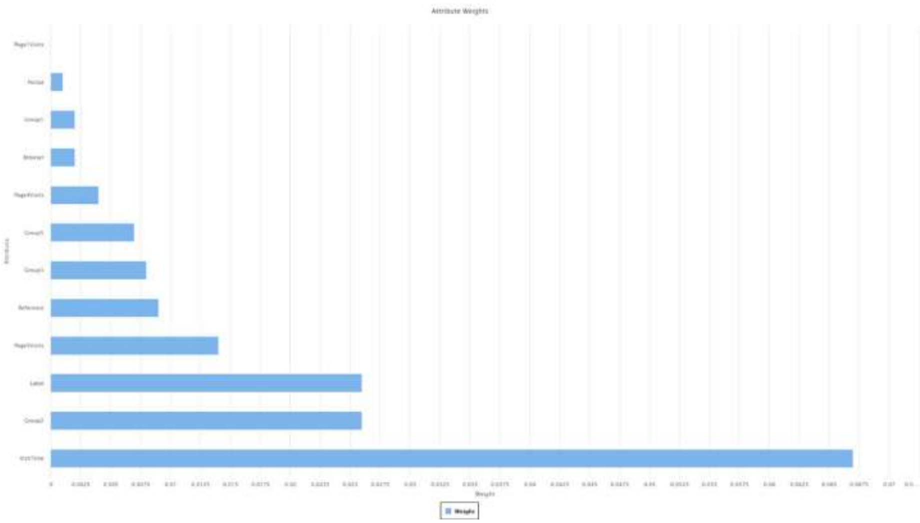

Figure 7. Presented the graphical view of weighted attributes

exchanging them. This technique is distinct from classic centralized machine learning models, that require all local dataset to be transmitted to a single server and much more traditional decentralized options, which often assume a uniform distribution of local data samples. Because it enables clients to cooperate on the train a global model using their data locally without exposing any with a third party, federated learning has sparked a lot of interest for data remote islands and privacy concerns. However, current federated learning frameworks usually need extensive condition setups (e.g. complex driver configuration of independent graphics cards such as NVIDIA, compilation environment), which are inconvenient to design and implement on a wide scale. We propose an innovative Web browser assist application that takes advantage of browser characteristics (e.g. Cross-platform, Java Programming language Characteristics) and improves privacy protection via a local differentially private mechanism to enable the implementation of federated learning or the integration of related applications. Finally, we run tests on a variety of devices to see how well the suggested Web analysis framework performs.

Model metric's

In the MSE, whatever is represented here on the vertical axis is plotted in units squared. The root mean squared error (RMSE) is another parameter that we compute (RMSE). It is the square root of the MSE. Because it is shown in similar units as the corresponding quantity, this statistic is likely to be the most comprehensible one. In Table 3, we represented the binomial GLM model, where the MSE obtained 0.016539 and RMSE obtained the 0.128604. In this table, RMSE obtained the better result than the MSE.

There are two types of labels as true and false. In this table, we predict the two labels. The true label obtained the 99.82% class precision, whereas the false label got the 0.47% class precision. The class recall is 97.66%of true false, whereas the true label class recall is 5.88%. Table 2 represented the performance vector result (Table 4).

	Generalized linear model	
Attributes	CO-EFF	STD CO-EFF
Label	6.3963058	6.3963058
Label.FT-MBA	4.403679868	4.403679868
Group2	2.546046633	1.543240765
Browser.Firefox	1.070265596	1.070265596
Reference.google	0.821594697	0.821594697
Label.ThoughtLeader	0.489867067	0.489867067
Reference.rotman	0.177344925	0.177344925
Group3	0.054478	0.449720511
Browser.InternetExplorer	0.041421131	0.041421131
Group1	0.01929874	0.123402972
Page4Visits	0.006119829	0.464525462
VistiTime	1.98×10^{-06}	0.627753826
Browser.BlackBerry	0	0
Page1Visits	-1.06×10^{-04}	−0.12580225
Browser.IEMobile	−0.003946981	−0.003946981
Label.InternationalStudent	−0.00603981	−0.00603981
Browser.Opera	−0.007005945	−0.007005945
Label.Creativity	−0.009207902	−0.009207902
Group5	−0.012172487	−1.817657953
Page5Visits	−0.02111315	−0.224336444
Browser.Unknown	−0.042586198	−0.042586198
Reference.linkedin	−0.044187911	−0.044187911
Reference.yahoo	−0.046853352	−0.046853352
Label.Finance-ThoughtLeadership	−0.193923019	−0.193923019
Label.Finance	−0.228798708	−0.228798708
Browser.Chrome	−0.249813291	−0.249813291
Reference.baidu	−0.407276756	−0.407276756
Label.Creativity-ThoughtLeadership	−0.457580853	−0.457580853
Reference.utoronto	−0.827565539	−0.827565539
Period.morning	−0.885844992	−0.885844992
Period.afternoon	−0.897312811	−0.897312811
Reference.other	−1.001136617	−1.001136617
Browser.Safari	−1.235560316	−1.235560316
Browser.Mozilla	−1.31551553	−1.31551553
Period.night	−1.585754549	−1.585754549
Browser.IE	−1.626170819	−1.626170819
Reference.bing	−2.040831798	−2.040831798
Intercept	−26.38455103	−8.717613185

Table 2. Existing the static model with generalized linear model

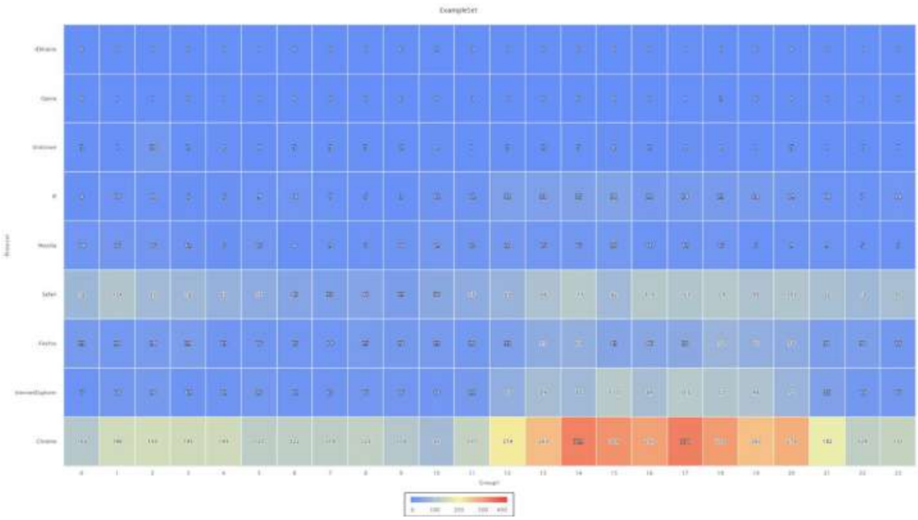

Figure 8. Display the static heat map of all groups

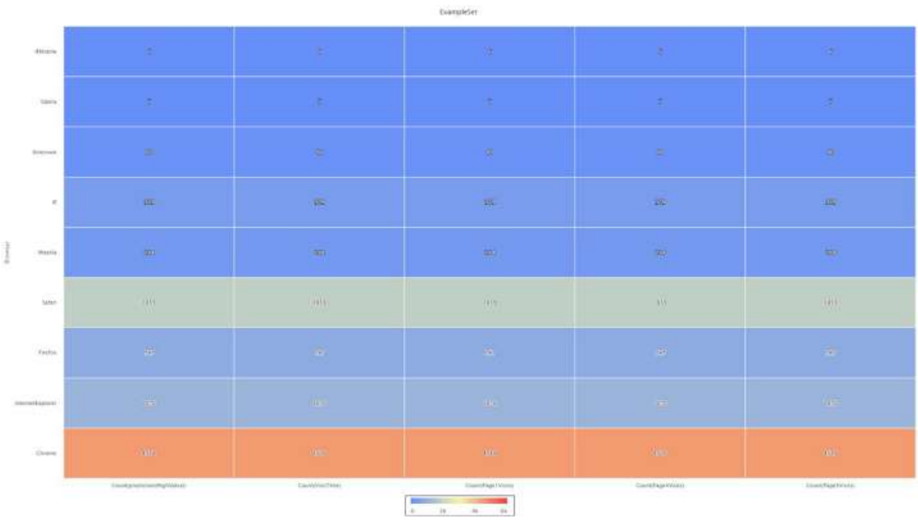

Figure 9. Displays the high value predict with a different browser

A classification problem's prediction outcomes are summarized in a matrix called a confusion matrix. There is a breakdown of the number of right and wrong predictions for each class based on the count values. This is the solution to the puzzle of the matrix of confusion. Confusion matrices are normalized over the actual (rows) and predicted (columns) circumstances or over the whole population. The matrix of confusion will not be normalized if none is selected. There are examples with the real label of Class I and the expected label of class j in a confusion matrix, which is represented by the *i*-th row and the *j*-th column entry. In Table 5, we predict the two labels. The true label error rate is 2 0.8824, whereas the false error rate is greater than true, which is 36 0.0120.

GLM families comprise a link function as well as a mean-variance relationship. For Poisson GLMs, the link function is a log, and the mean–variance relationship is the identity. As for the link function, it allows us to model nonlinear relationships between our predictors and our response. In a simple linear regression our model the expected value directly as a linear combination of the predictors. In a GLM, on the other hand, our model is a function of the expected value. Table 6 presented the GLM model summary family link.

In Table 7, we described the scoring history of 20 iterations. Every iteration has different values negative_log_likelihood and objective, where the 0 iteration negative_log_likelihood is 104.9917 and objective is 0.0348. However, we draw more iterations for negative_log_likelihood. The value is decreased with each iteration like

Table 3. Presented the binomial GLM model

Model metrics	
Model ID	Binomial GLM
MSE	0.016539
RMSE	0.128604
R^2	1.951816
AUC:	0.771382
pr_auc	0.019696
logloss	0.078329
mean_per_class_error	0.447176
default threshold	0.455135

Table 4. Presented the performance vector of the predicted class

Performance vector			
Predict label	True false (%)	True true (%)	Class precision (%)
pred. false	8,773	16	99.82
pred. true	210	1	0.47
class recall	97.66	5.8	

Table 5. Presented the confusion matrix

CM: confusion matrix			
Row labels	Actual class	Column labels	Predicted class
False	True	Error	Rate
False	2,964	36 0.0120	36/3,000
True	15	2 0.8824	15/17
Totals	2,979	38 0.0169	51/3,017

Table 6. Presented the GLM model summary family link

GLM model (summary)			
Family Link	Regularization no. of predictors	Total no. of active predictors	No. of iterations
binomial logit Ridge (lambda = 8.607×10^{-6})	37	36	20

Iterations	Scoring history of iterations Negative_log_likelihood	Objective
0	104.9917	0.0348
1	96.95322	0.03214
2	85.65288	0.0284
3	82.18758	0.02726
4	75.17974	0.02498
5	72.9251	0.02424
6	68.49899	0.02283
7	64.75808	0.02164
8	63.24154	0.02112
9	61.99785	0.02073
10	61.06093	0.02046
11	60.09731	0.02017
12	59.65219	0.02007
13	59.41097	0.01999
14	58.70673	0.01978
15	58.49177	0.01972
16	58.26448	0.01965
17	58.00651	0.01957
18	57.91105	0.01952
19	57.64456	0.01946
20	57.46807	0.01942

Table 7. Presented the iteration scores

the final iteration 20 for negative_log_likelihood is 57.46807. As well as the objective is also reduced with each iteration. Iteration 20 objective value is 0.01942.

Conclusion

Federated learning has emerged as a training paradigm in such settings. Web browser forensics research in general has focused on certain browsers or the computational modeling of specific log files. Internet users engage in a wide range of activities using an internet browser, such as searching for information and sending e-mails. It is also essential that the investigator have access to user activity when conducting an inquiry. This data, which may be used to assess information retrieval activities, is very critical. In this paper, we purposed a major tool used for Web browser analysis. Our proposed algorithm is capable of protecting data privacy effectively in real-world experiments.

Future work and limitations

Web browser forensics will be studied in the future under a variety of operating systems, including Windows, Mac, Laptop and mobile devices. This work is limited to 9,000 sample data set, with 16 attributes. In the future Authors can work on the big data with maximum attributes.

References

Akanksha, E., Jyoti, N.S. and Gulati, K. (2021), "Review on reinforcement learning, research evolution and scope of application", *Proceedings – 5th International Conference on Computing Methodologies and Communication, ICCMC 2021*, doi: 10.1109/ICCMC51019.2021.9418283.

Akanksha, E., Sharma, N. and Gulati, K. (2021), "OPNN: Optimized probabilistic neural network based automatic detection of maize plant disease detection", *Proceedings of the 6th International Conference on Inventive Computation Technologies, ICICT 2021*, doi: 10.1109/ICICT50816.2021.9358763.

Akanksha, E., Sharma, N. and Gulati, K. (2021), "Review on reinforcement learning, research evolution and scope of application", *2021 5th International Conference on Computing Methodologies and Communication (ICCMC), IEEE*, pp. 1416-1423.

Akanksha, E., Sharma, N. and Gulati, K. (2021), "OPNN: optimized probabilistic neural network based automatic detection of maize plant disease detection", *2021 6th International Conference on Inventive Computation Technologies (ICICT), IEEE*, pp. 1322-1328.

Alam, F., Rahman, S.U., Ullah, S. and Gulati, K. (2018), "Medical image registration in image guided surgery: issues, challenges and research opportunities", *Biocybernetics and Biomedical Engineering*, Vol. 38 No. 1, doi: 10.1016/j.bbe.2017.10.001.

Bansal, S., Gulati, K., Kumar, P. and Choudhury, T. (2018), "An analytical review of PaaS-cloud layer for application design", *Proceedings of the 2017 International Conference On Smart Technology for Smart Nation, SmartTechCon 2017*, doi: 10.1109/SmartTechCon.2017.8358374.

Billewar, S.R., Jadhav, K., Sriram, V.P., Arun, D.A., Mohd Abdul, S., Gulati, K. and Bhasin, D. (2021), "The rise of 3D E-Commerce: the online shopping gets real with virtual reality and augmented reality during COVID-19", *World Journal of Engineering*, Vol. 19 No. 2, doi: 10.1108/WJE-06-2021-0338.

Dhiman, G., Juneja, S., Viriyasitavat, W., Mohafez, H., Hadizadeh, M., Islam, M.A., El Bayoumy, I. and Gulati, K. (2022), "A novel machine-learning-based hybrid CNN model for tumor identification in medical image processing", *Sustainability*, Vol. 14 No. 3, p. 1447, doi: 10.3390/su14031447.

Dovhan, O.D., Yurchenko, O.M., Naidon, J.O., Peliukh, O.S., Tkachuk, N.I. and Gulati, K. (2021), "Formation of the counter intelligence strategy of Ukraine: national and legal dimension", *World Journal of Engineering*, Vol. 19 No. 2, doi: 10.1108/WJE-06-2021-0358.

Gulati, K. and Telu, P. (2016), "A study of progress of E-Participation in India", *International Conference on Electrical, Electronics, and Optimization Techniques, ICEEOT 2016*, doi: 10.1109/ICEEOT.2016.7755360.

Gulati, K., Boddu, R.S.K., Kapila, D., Bangare, S.L., Chandnani, N. and Saravanan, G. (2021a), "A review paper on wireless sensor network techniques in internet of things (IoT)", in *Materials Today: Proceedings, Publication Elsevier 2021*, Vol. 51, ISSN 2214-7853, doi: 10.1016/j.matpr.2021.05.067, 19 May 2021.

Gulati, K., Kumar, S.S., Boddu, R.S.K., Sarvakar, K., Sharma, D.K. and Nomani, M.Z.M. (2021b), "Comparative analysis of machine learning-based classification models using sentiment classification of tweets related to COVID-19 pandemic", in *Materials Today: Proceedings, 2021*, ISSN 2214-7853, doi: 10.1016/j.matpr.2021.04.364, 12 May 2021.

Joseph, L.M.I.L., Goel, P., Jain, A., Rajyalakshmi, K., Gulati, K. and Singh, P. (2021), "A novel hybrid deep learning algorithm for smart city traffic congestion predictions", *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, pp. 561-565, doi: 10.1109/ISPCC53510.2021.9609467.

Sanil, H.S., Singh, D., Raj, K.B., Choubey, S., Bhasin, N.K.K., Yadav, R. and Gulati, K. (2021), "Role of machine learning in changing social and business eco-system – a qualitative study to explore the factors contributing to competitive advantage during COVID pandemic", *World Journal of Engineering*, Vol. 19 No. 2, doi: 10.1108/WJE-06-2021-0357.

Sriram, V.P., Raj, K.B., Srinivas, K., Pallathadka, H., Sajja, G.S. and Gulati, K. (2021), "An extensive systematic review of RFID technology role in supply chain management (SCM)", *2021 6th International Conference on Signal Processing, Computing and Control (ISPCC)*, pp. 789-794, doi: 10.1109/ISPCC53510.2021.9609414.

Virtanen, J.P., Jaalama, K., Puustinen, T., Julin, A., Hyyppä, J. and Hyyppä, H. (2021), "Near real-time semantic view analysis of 3D city models in web browser", *ISPRS International Journal of Geo-Information*, Vol. 10 No. 3, doi: 10.3390/ijgi10030138.

Wassan, S. (2021), "How artificial intelligence transforms the experience of employees", *Turkish Journal of Computer and Mathematics Education*, doi: 10.17762/turcomat.v12i10.5603.

Wassan, S., Shen, T., Xi, C., Gulati, K., Vasan, D. and Suhail, B. (2022), "Customer experience towards the product during a coronavirus outbreak", *Behavioural Neurology*, Vol. 2022, p. 18, doi: 10.1155/2022/4279346. Article.ID 4279346.

Wisetsri, W., Julie Aarthy, R.T.S.C.C., Thakur, V., Pandey, D. and Gulati, K. (2021), "Systematic analysis and future research directions in artificial intelligence for marketing", in *Turkish Journal of Computer and Mathematics Education (SCOPUS INDEXED)*, Vol. 12 No. 11, pp. 43-55, e-ISSN 1309-4653, doi: 10.17762/turcomat.v12i11.5825, 10 May 2021.

Zhioua, S. (2015), "The web browser factor in traffic analysis attacks", *Secur. Commun. Networks*, doi: 10.1002/sec.1338.

Corresponding author
Kamal Gulati can be contacted at: drkamalgulati@gmail.com