# Not all web pages are born the same content tailored learning for web QoE inference

**Pedro Casas, Sarah Wassermann, Nikolas Wehner, Michael Seufert, Tobias Hoßfeld**

**Angaben zur Veröffentlichung / Publication details:**

Casas, Pedro, Sarah Wassermann, Nikolas Wehner, Michael Seufert, and Tobias Hoßfeld. 2022. "Not all web pages are born the same content tailored learning for web QoE inference." In *IEEE International Symposium on Measurements & Networking (M&N), 18-20 July 2022, Padua, Italy*, edited by Marco Carratu', 1–6. Piscataway, NJ: IEEE. https://doi.org/10.1109/mn55117.2022.9887781.



# Not all Web Pages are Born the Same Content Tailored Learning for Web QoE Inference

Pedro Casas*, Sarah Wassermann*, Nikolas Wehner†
Michael Seufert†, Tobias Hossfeld†
*AIT Austrian Institute of Technology, †University of Würzburg

***Abstract*—Web Quality of Experience (QoE) monitoring is a critical task for Internet Service Providers (ISPs), especially due to the key role played by customer experience in churn management. Previously, we have tackled the problem of Web QoE inference from the ISP perspective, relying on passive measurement of encrypted network traffic and machine learning models. In this paper, we exploit the broad heterogeneity of contents embedded in web pages to improve the state of the art performance in Web QoE inference, relying on web-content learning model tailoring. By analyzing the top-500 most popular web pages of the Internet through unsupervised learning, we discover different web page content classes which realize significantly different Web QoE inference performance. We train supervised learning inference models separately for each of these classes, using the well-known Speed Index (SI) metric as proxy to Web QoE. Empirical evaluations on a large corpus of Web QoE measurements for top popular websites demonstrate that our combined content-tailored approach improves the inference performance of the SI by almost 30% with respect to previous single-model approaches, reducing the QoE inference error in terms of mean opinion scores by more than 40%.**

## I. Introduction

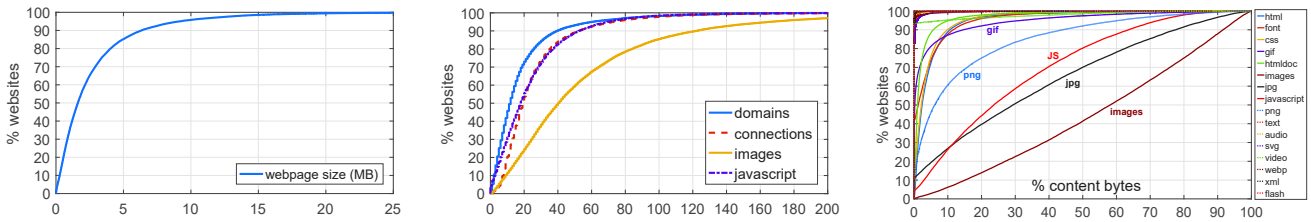
Web browsing Quality of Experience (QoE) has attracted significant attention in recent years. While Internet Service Providers (ISPs) have traditionally relied on the usage of Deep Packet Inspection (DPI) techniques to understand the performance of web services from the network side, the wide adoption of end-to-end traffic encryption has drastically reduced their visibility. This has motivated a surge in the research and conception of Machine-Learning (ML) based approaches to infer application-level Web QoE metrics from the streams of encrypted bytes [4], [6]. In these previous work, single learning models were trained to address the full spectrum of web pages in the Internet. Given the huge number of Internet web pages – more than 1.7 billion websites (https://httparchive.org/), and the diversity of contents embedded on them, we hypothesize that a single model cannot efficiently capture this rich heterogeneity. While it is unfeasible to build per-web page specific models, we investigate to which extent we can identify groups of web pages sharing similarities in their underlying structure (e.g., image-dominant vs text-dominant pages, number of external embedded contents, small vs large page size, etc.), to train per-group inference models which realize better overall performance. Previous studies [9], [10] have suggested this is the case, but no actual research agenda was followed-up on this problem. In a more general perspective, this problem is related to the personalization of machine learning models [13]: in machine learning, personalization addresses the goal of training a model to target a particular individual or homogenous group, significantly enhancing the realized performance.

In this paper we build on our previous work [4], [5] to address this personalization approach, firstly by discovering classes of web pages sharing similar content-related properties through clustering techniques, and then by training per-class supervised learning models to infer the SI of individual web page loading sessions. We take the well-known SI metric as a proxy to Web QoE, based on the rich literature on Web QoE analysis [4]–[9]. To do so, we analyze a rich dataset of active Web QoE measurements, targeting the most popular websites in today's Internet. The dataset includes both application-layer Web QoE metrics – such as SI, as well as network traffic traces, for 15,000 web page *loading sessions* (i.e., the loading of a single browser web page). Finally, while we have recently worked on the problem of clustering web pages based on content characteristics [3], this paper goes beyond our previous work and the state of the art, by enhancing Web QoE inference through content-tailored machine learning. In particular, we show that our combined, content-tailored approach improves the inference performance of the SI by almost 30% with respect to previous one-fit-all single model approach, additionally reducing the QoE inference error in terms of Mean Opinion Scores (MOS) by more than 40%.

The remainder of the paper is organized as follows. Sec. II overviews the related work on Web QoE monitoring and analysis. Sec. III presents the overall modeling and inference approaches as well as the data generation, including a description of the different features used in the unsupervised and supervised tasks. In Sec. IV we apply clustering algorithms to automatically discover different classes of web pages based on their contents, and present a characterization of the obtained results. Using these classes as basis, in Sec. V we benchmark and explain the performance of per-class supervised learning models for SI inference against the previously followed one-fit-all model. Finally, Sec. VI concludes this paper.

## II. Related Work

There is a vast literature in the problem of measurements for Web QoE monitoring and analysis [2], [7]–[9], [11]; however, most of previous work have focused on measurements at the application layer or assuming access to end devices, which is not applicable for ISPs to perform network-wide QoE




(a) Webpage size (MB). (b) Number of domains, connections, and relevant resources. (c) Share of content types.

Fig. 1: Heterogeneity of Internet web pages, for the top 10.000 Alexa websites in 2021.

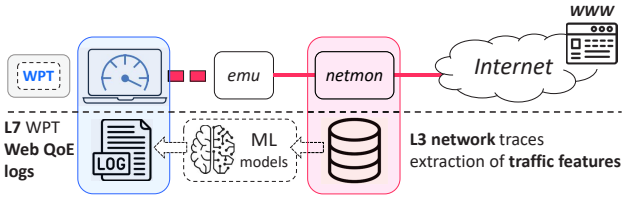



Fig. 2: Measurement platform for data collection.

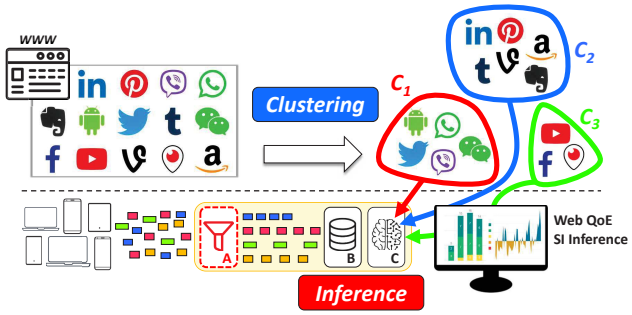



Fig. 3: Web page classes discovery and SI inference.

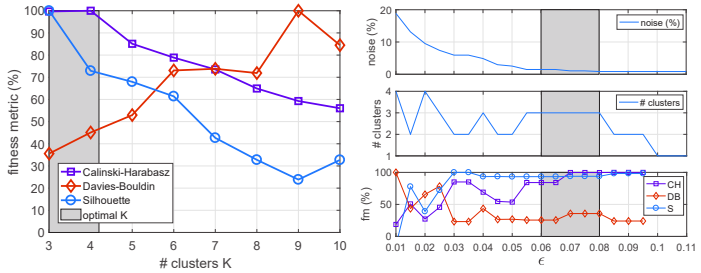
monitoring in practice. ISPs require approaches which can operate directly at the network traffic level, where massive data is available to their already deployed monitoring systems, yet with the additional complexity introduced by the wide adoption of TLS/HTTPS for end-to-end traffic encryption. Previous work [1], [2] have developed Web QoE-related metrics highly correlated to the SI metric, including the Byte and Object-Index [2] and the Pain-Index [1], which can be computed directly from packet and flow level measurements, thus seamlessly operating with encrypted traffic. Still, such metrics are mostly informative, as they do not provide an absolute estimation of the actual user QoE. The SI metric is today widely accepted as one of the best metrics serving as proxy to Web QoE [7], as it takes into account the whole visual progress of the page loading. As a consequence, and considering the monitoring limitations introduced by network traffic encryption, in recent work [4]–[6] we have tackled the problem by conceiving ML models to directly infer the SI of web loading sessions, using inputs directly derived from the encrypted streams of traffic. In this paper we take a step forward in this direction, by conceiving models for classes of web pages sharing similar content characteristics, which eventually results in an enhanced performance of the SI inference task. Explainability analysis through well-known approaches such as SHAP [12] show that simple multi-flow-level metrics such as session duration and flow index (reflecting download throughput) provide the most descriptive information behind such an improved performance.

## III. Web QoE Data & Modeling Approaches

To show evidence on the broad heterogeneity and content richness of Internet web pages today, Fig. 1 depicts the (a) size, (b) number of domains and relevant resources, and (c) share of content types, for the top 10.000 Alexa websites in 2021, extracted from the HTTP Archive public dataset (https://httparchive.org/) in Google BigQuery.

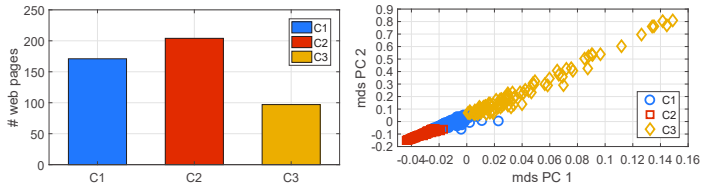
The proposed solution to the Web QoE inference problem consists of training supervised ML models to map network traffic features, extracted from the encrypted network web-page traffic, into the SI metric. The approach is data-driven, and thus needs datasets containing both the collected network traffic – the *input*, and the targeted Web QoE metric – the *ground truth*. To fully control the generation of such datasets, we have conceived a measurement platform based on multiple private instances of WepPageTest (WPT) [14], a well-known and widely used open-source web performance analysis tool. The measurement testbed (Fig. 2) consists of three different, non-emulated types of devices, including smartphones, tablets, and desktop (Chrome is used as browser), using WPT agents for Android and Linux. To keep the focus on the web page content aspect solely, we use only measurements collected for Desktop browsing in this study. Using WPT measurements, the platform extracts about 90 different KPIs and Web QoE metrics from independent web page loading sessions – including the SI metric, as well as content characteristics of the visited web pages. Network traffic is captured at an intermediate monitoring vantage point, from where different traffic features are subsequently derived as input to the models. A desktop device is connected to the open Internet through a network emulator; we use different configurations including downlink bandwidth up to 10 Mbps, packet loss up to 10%, and RTTs up to 100ms. Web measurements target the top 500 most popular websites according to Alexa top sites list. The same web pages are visited multiple times, under the same access network setups. For this study, we use a dataset composed of 15,000 individual web page loading sessions.

Using these measurements, the proposed modeling approach is two-fold (Fig. 3): the first step (top) consists of the identification of web page content classes $C_i$, which is




(a) $K$-means. (b) DBSCAN.

Fig. 4: Identification of web page clusters.




(a) # wepages per cluster. (b) Multidimensional scaling.

Fig. 5: Obtained web page clusters with $K = 3$-means.

portrayed as an unsupervised learning task, more precisely relying on clustering techniques. To tackle this problem, we define a set of content-level features $F_C = \{f_i\}$ describing the characteristics of each web page in terms of page size, number of embedded contents, root domains, and share of bytes for different MIME content types (e.g., images, video, javascript, etc.). Tab. I presents the set of 16 different content features used for web pages clustering. This clustering step is independent of the SI inference problem, and relies on content-related features which are not necessarily visible at the network traffic level.

The second step (down) consists of the training of supervised learning models (cf. module C in Fig. 3) which can infer the SI of individual web page loading sessions from features derived out of the (encrypted) network traffic flows (cf. module B in Fig. 3). To compute these features, we assume that all the traffic flows belonging to an individual web page loading session are already isolated from the rest of the traffic, which in the practice would be done by a monitoring system in module A. While this traffic monitoring module is out of the scope of the paper, we acknowledge that we have conceived a fingerprinting solution to identify all the traffic flows belonging to a specific web page, relying on DNS information. For the SI inference problem, we take the set of 21 flow-level features previously used in [4], including: $(i)$ the total number of flows (all, downlink, uplink), $(ii)$ the min/mean/median/max flow duration in downlink, $(iii)$ the min/mean/median/max flow size in downlink, $(iv)$ the min/mean/median/max flow Byte Index [2] in downlink, $(v)$ the mean/median in-flow, average intra-packets time (MDT) in downlink, $(vi)$ the mean/median/max flow throughput in downlink, and $(vii)$ the Flow Index (FI). The FI metric [4] is the equivalent to the Byte Index, counting downloaded flows instead of bytes. These flow-level features are complemented by a set of 11 session-level features (computable at the flow-level), which include: $(i)$ session duration

| ID | feature | ID | feature |
|---|---|---|---|
| $f_1$ | # $\text{packets}_{\text{down}}$ | $f_9$ | # requests HTML |
| $f_2$ | # $\text{packets}_{\text{up}}$ | $f_{10}$ | % $\text{bytes}_{\text{html}}$ |
| $f_3$ | # packets | $f_{11}$ | % $\text{bytes}_{\text{js}}$ |
| $f_4$ | # $\text{bytes}_{\text{down}}$ | $f_{12}$ | % $\text{bytes}_{\text{css}}$ |
| $f_5$ | # $\text{bytes}_{\text{up}}$ | $f_{13}$ | % $\text{bytes}_{\text{img}}$ |
| $f_6$ | # bytes | $f_{14}$ | % $\text{bytes}_{\text{flash}}$ |
| $f_7$ | # root domains | $f_{15}$ | % $\text{bytes}_{\text{font}}$ |
| $f_8$ | # resources | $f_{16}$ | % $\text{bytes}_{\text{video}}$ |

TABLE I: Features $F_C$ for web pages clustering.

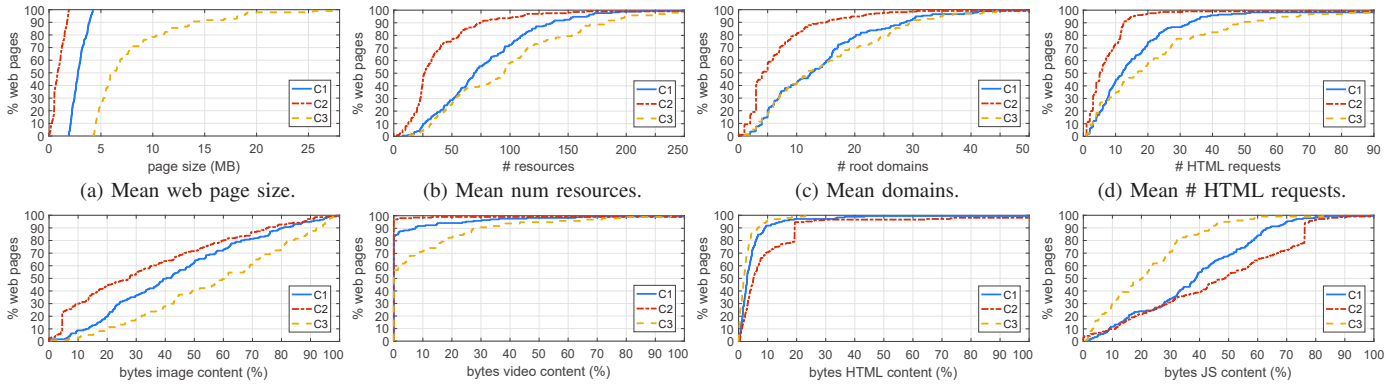
(total, downlink, uplink), $(ii)$ total number of packets and bytes (all, downlink, uplink), and $(iii)$ mean session throughput in downlink and uplink. We refer to this set of 32 network traffic features as $F_N$. Finally, the content-tailored modeling is realized by training a separate SI inference model $i$ for all the measurements corresponding to the web pages in $C_i$.

## IV. Clustering Web Pages

To split the set of $n = 500$ web pages into meaningful clusters, we firstly compute the set of 16 $F_C$ features, out of the 15,000 web page loading sessions available in the dataset. For each web page $WP_i$, $i = 1..n$, features $f_j(i)$, $j = 1..9$ are computed as the mean value observed over the multiple loading sessions of this web page, and features $f_j(i)$, $j = 10..16$ are computed as the shares of MIME content types for the average web page $\overline{WP}_i$. The effective number of analyzed web pages is $n = 472$ and not $n = 500$, which is due to repeatedly timed-out measurements for web pages located mostly in China.

The next step is to decide on the specific clustering approach to apply. The two most well-known and standard clustering algorithms are $K$-means and DBSCAN. The former identifies a pre-defined number of clusters $K$, with hyper-spherical shape when considering a dimension-independent distance metric such as the Euclidean distance; the latter works with the concepts of density and outliers, and identifies any shape of cluster containing at least $min_{pts}$ samples, using the distance $\epsilon$ as basis for density-search. In our particular problem, a practical advantage of $K$-means over DBSCAN is that each instance is always assigned to a cluster, which is not the case for DBSCAN, which labels instances as outliers (noise) when located in low-density regions with less than $min_{pts}$ samples in an $\epsilon$-vicinity. We therefore consider $K$-means as the clustering algorithm to use. To find the optimal number of clusters $K$ we take a combined grid-search approach, using both $K$-means and DBSCAN as search algorithms, and using a set of structural performance metrics to assess the obtained clustering results when changing either $K$ or $\epsilon$. In particular, we use the well known variance ratio criterion – Calinski-Harabasz (CH) index, the DaviesBouldin (DB) index, and the Silhouette (S) index. We refer the reader to [15] for a detailed definition of these cluster validity metrics, but in a nutshell, the higher the CH and S indexes, and the lower the DB index, the better the clustering results.

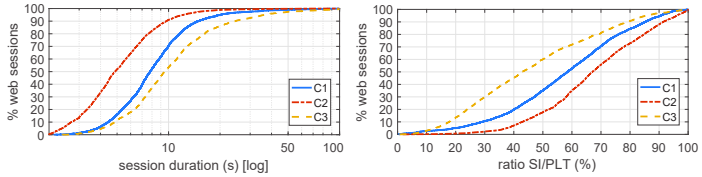
Fig. 4 depicts the obtained results using (a) $K$-means with a growing number of clusters, and (b) DBSCAN with a growing




(a) Mean web page size. (b) Mean num resources. (c) Mean domains. (d) Mean # HTML requests.

(e) % IMG bytes. (f) % video bytes. (g) % HTML bytes. (h) % JS bytes.

Fig. 6: Characterization of web pages per cluster. Web page size is a clear differentiator for cluster membership.




(a) Loading session duration. (b) SI/PLT ratio.

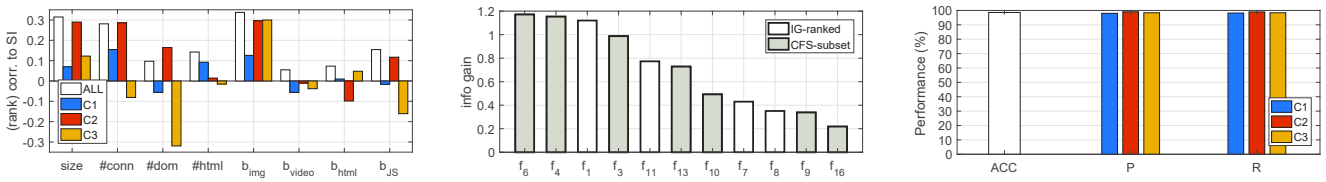
Fig. 7: Web page loading performance and complexity.

search distance $\epsilon$, fixing $min_{pts} = 5$; note that the influence of $min_{pts}$ is generally much lower [15]. According to (a), $K = 3$ or $K = 4$ provides the best clustering performance in terms of the validity metrics; to take a more informed decision, we rely on the results obtained with DBSCAN, not only in terms of the same validity metrics, but also considering the number of identified clusters and the corresponding percentage of outliers. The number of clusters identified by DBSCAN significantly fluctuates between two and four, but a certain stability is observed for $0.055 \leq \epsilon \leq 0.08$, resulting in three identified clusters with high performance and a very low share of outliers. Therefore, the final number of clusters to consider is set to $K = 3$. Fig. 5 shows (a) the number of web pages identified within each cluster $C_i$, as well as (b) the similarity among web pages, through a multidimensional scaling approach. The size of each cluster is $|C_1| = 171$, $|C_2| = 204$, and $|C_3| = 97$, and their split has limited overlapping; as we show next, the web page size is a highly accurate feature to differentiate web pages.

### A. Characterization of Web page Classes

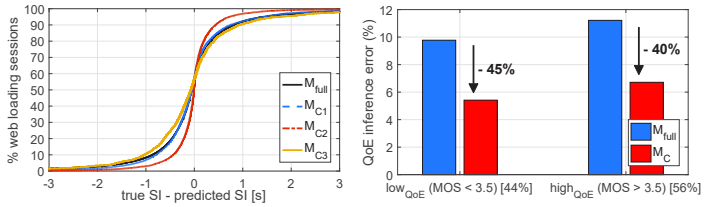
Fig. 6 depicts different characteristics of the web pages, split by cluster. The most relevant observation relates to the influence of the web page size: according to Fig. 6(a), $C_2$ contains smaller-size web pages of up to 2MB, $C_1$ contains web pages with sizes between 2MB and 4MB, and $C_3$ contains a broader spectrum of web pages, with size ranging from about 4.2MB up to 30MB. This ordering is maintained when considering the number of embedded resources and external domains, as well as the number of required HTML requests to load the page. In terms of MIME content types, while similar observations apply – e.g., $C_3$ web pages are characterized by a higher prevalence of image and video contents, $C_2$ web pages have more HTML and javascript content, suggesting a more interactive nature of these web pages. In terms of loading performance, Fig. 7 shows that $C_1$ and $C_3$ web pages take significantly higher time to fully download all required contents as compared to $C_2$ web pages, which is expected based on the aforementioned content observations. Interestingly, the SI to page loading time (PLT) ratio as depicted in Fig. 7(b) evidences that the higher the size and complexity of a web page, the bigger the difference between the visually perceived loading time – i.e., the SI, and the total loading time.

We investigate the correlation of some of the considered content features to the SI, as well as the relevance of the different features in terms of defining the clustering membership of a web page. Fig. 8(a) depicts rank correlation values between web page contents and SI, for all web pages and in a per-cluster basis. While correlation values are rather low in most cases, $C_2$ web pages loading performance is more correlated to the underlying contents. Page size, number of resources, and share of image bytes are among the most relevant features to infer the SI for web pages in $C_2$. Fig. 8(b) ranks the set of 16 features according to the information gain provided by each feature, when using the features as input for per-cluster content classification. In simple terms, the higher the information gain, the more relevant the feature is to discriminate among clusters. We also flag subsets of features providing the highest correlation values to the cluster membership, through correlation feature selection techniques. As expected, page-size related features are ranked first, including page size ($f_6$) and downloaded bytes ($f_4$) and packets ($f_1$); image ($f_{13}$) and JS ($f_{11}$) contents are also relevant for web page discrimination, as well as the number of external root domains ($f_7$) and embedded resources ($f_8$). Finally, as an additional verification of the goodness of clustering results, Fig. 8(c) reports the cross-validated web-page-to-cluster identification performance, assuming here a random forest (RF) model as supervised classifier, and the clustering membership as ground-truth. In simple terms, this model could accurately associate a web page to its corresponding category, if the corresponding content-related features $F_C$ would be exposed.




(a) Correlation to SI. (b) Feature ranking for cluster identification. (c) Cluster identification performance.

Fig. 8: Correlation of $F_C$ features to SI and cluster membership.




(a) Distribution of inference errors. (b) True vs. predicted SI.

Fig. 9: SI inference performance.

| model (content) | MAE-mAE (ms) | MRE-mRE (%) | PLCC |
|---|---|---|---|
| full | 660 – 303 | 35 – 18 | 0.847 |
| $C_1$ | 540 – 285 | 28 – 15 | 0.906 |
| $C_2$ | 308 – 141 | 24 – 11 | 0.942 |
| $C_3$ | 712 – 382 | 33 – 17 | 0.890 |
| $C$ | 474 – 223 | 27 – 13 | 0.916 |

TABLE II: SI inference by content-tailored ML.

## V. Content-Tailored Web QoE Inference

### A. Web QoE Inference through Tailored Models

We now take the three identified clusters of web pages to train independent SI inference models in a per-cluster basis. For each cluster $C_i$, we take a subset of the 15,000 web page loading sessions, corresponding to all the loading sessions of the member web pages. In practice, the association of a web page to its corresponding cluster class can be simply done by inspection of the underlying DNS request.

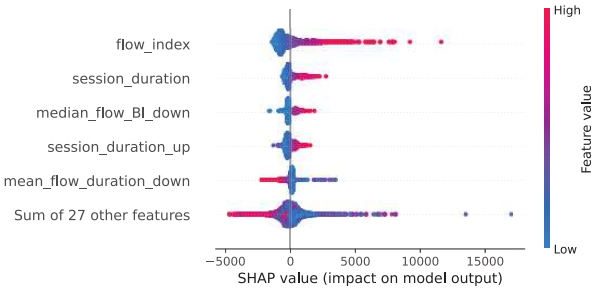
Following our previous studies on ML for Web QoE [4], we take a simple RF model to tackle the inference task. We consider an expressive model composed of 100 decision trees, which is trained and evaluated through cross-validation. More specifically, all results presented next correspond to 5-fold cross validation. Recall that in the SI inference problem, we take as input the set of 32 features $F_N$ computed at the flow level, directly from the encrypted stream of traffic.

To assess the relevance of the identified web page clusters regarding the improvement of the inference models, we compare the SI inference performance realized by a single model cross-validated over the full set of web page measurements – we refer to this model as $\mathrm{M}_{\mathrm{full}}$, against three independent $\mathrm{M}_{C_i}$ models, cross validated over member web page measurements. We assess performance using three standard performance metrics for regression problems, including the absolute error (AE), the relative error (RE), and the linear correlation (PLCC). We take both mean (M) and median (m) values for the error metrics, to filter out significantly large errors.

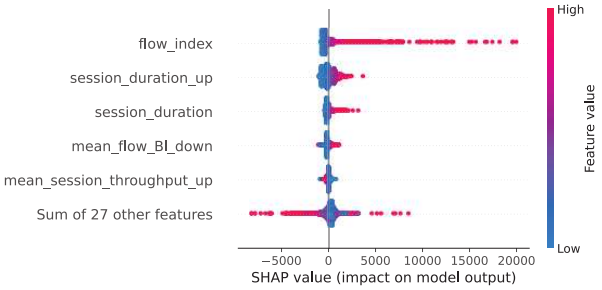
Fig. 9(a) depicts the distribution of the inference errors for the aforementioned models. A first observation is that the $\mathrm{M}_{C_2}$ model largely outperforms the others, providing significantly lower errors. This is coherent with previous observations on the correlation of content-related metrics and SI for $C_2$ web pages, and in particular to page size related metrics, which are easily capture also at the flow-level measurements. Inference performance for $\mathrm{M}_{C_1}$ is also better than for $\mathrm{M}_{\mathrm{full}}$, and $\mathrm{M}_{C_3}$ realizes worse results. Tab. II summarizes the obtained results. In particular, the last row indicated by $C$ corresponds to all clustering prediction results realized by the three models $\mathrm{M}_{C_i}$, each applied to the corresponding loading sessions for the web pages within $C_i$. In terms of absolute inference errors, the content-tailored inference approach improves the single $\mathrm{M}_{\mathrm{full}}$ model by almost 30%, reducing the mean absolute error from 660ms to 474ms. A similar improvement is observed for the median absolute error, reducing from 303ms to 223ms. A natural question here is how relevant are these improvements in terms of end-user experience? To answer this question, we map SI values into well-known MOS scores, using waiting-time models calibrated in [7]. For the sake of easier analysis, we map SI values into low and high QoE categories, being low/high QoE a MOS score below/above 3.5. This threshold also results in a balanced split (44/56) in terms of number of sessions in each category. Fig. 9(b) depicts the corresponding QoE MOS inference errors obtained when mapping actual and inferred SI values, for $\mathrm{M}_C$ aggregated per-cluster models, and for $\mathrm{M}_{\mathrm{full}}$. $\mathrm{M}_{\mathrm{full}}$ inference errors result in about 10%/11% of the sessions being wrongly classified as low/high QoE. Aggregated $\mathrm{M}_C$ results reduce these errors by more than 40%, improving low/high QoE classification by 45%/40%.

### B. Model Explainability with SHAP

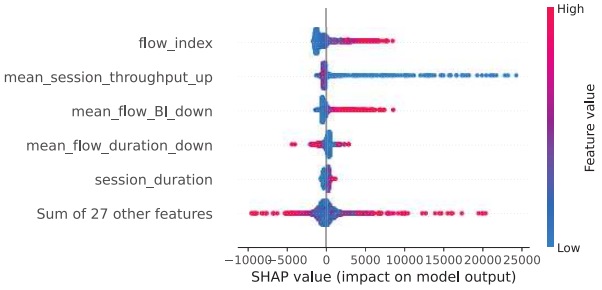
To shed light on the underlying functioning of the RF models, we resort to explainable AI techniques describing which input features have a stronger influence in the predictions, and in which sense. We apply SHAP [12], a game theoretic approach to explain the output of any machine learning model, by calculating the contribution of each feature to the corresponding predictions. Fig. 10 depicts the top-5-sorted most relevant input features and their corresponding impact on the model output according to SHAP, per cluster model. For each feature and for each input sample, each plot shows how much it pushes the model output from the base value to the actual output for this sample. The base value corresponds to the average model output over the training dataset. Higher feature values are shown in red, and lower




(a) SHAP top-5 features for $M_{C_1}$ and $C_1$ web pages.




(b) SHAP top-5 features for $M_{C_2}$ and $C_2$ web pages.



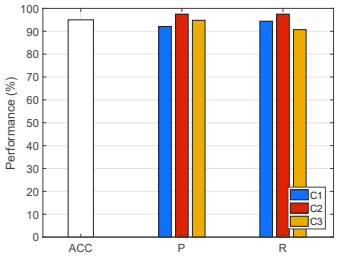

(c) SHAP top-5 features for $M_{C_3}$ and $C_3$ web pages.

Fig. 10: Explainability of per-cluster SI inference models.

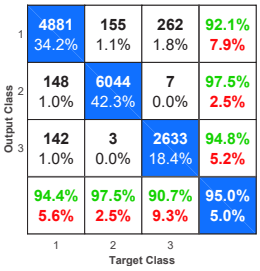
in blue. Multi-flow-level features such as flow index and mean/median flow byte index (for both metrics, the smaller the value, the faster the page loading and the smaller the SI), and session duration have a clearer stronger impact, positively correlated with the SI. Interestingly, the uplink throughput is highly relevant for the $M_{C_3}$ model; web pages in cluster $C_3$ are in general more complex (cf. Fig. 6), with a higher number of resources and images, and requiring more HTML requests for content-fetching, which means more round-trip-times, and therefore a higher dependence on bi-directional link performance. A deeper analysis on the impact of the uplink performance is part of ongoing work.

### C. Classification of Web Pages

The last part of the study is devoted to the problem of automatically assigning web page loading sessions to the corresponding web page class when no DNS information is available. This could happen for multiple reasons, and thus it is relevant to understand how feasible it would be to predict the class membership of a web page using flow-level traffic features only, to apply the corresponding content-tailored inference model. Fig. 11 shows that web pages can be




(a) Classification performance.




(b) Confusion matrix.

Fig. 11: Classification of web pages from flow-metrics.

properly assigned to their underlying classes with high precision and recall – above 90% for all three classes, exclusively from network traffic measurements and even when no DNS information is available.

## VI. Concluding Remarks

The broad heterogeneity of contents embedded in modern web pages can be exploited to provide more accurate models for Web QoE inference, through personalization and tailoring of machine learning models. We have shown that content-tailored models can significantly enhance the performance of previous machine learning driven solutions for in-network Web QoE monitoring, providing the first tangible evidence of this potential, for a yet not answered question. We discovered that the top-500 most popular web pages of the Internet can be grouped under three different content-based classes, and that the size of a page alone provides already relevant enough information for a raw split of web pages into meaningful content classes. In future work, we plan to extend the clustering of web pages and tailoring of machine learning models to a much larger set of Internet web pages, evaluating the benefits and limits of content-tailored learning for Web QoE monitoring and assessment at large.

## References

[1] M. Trevisan et al., "PAIN: A Passive Web Performance Indicator for ISPs," *Computer Networks*, vol. 149, 2019.
[2] E. Bocchi et al., "Measuring the Quality of Experience of Web Users," *ACM SIGCOMM CCR*, vol. 46(4), 2016.
[3] L. Jiménez et al., "Content Matters: Clustering Web Pages for QoE Analysis with WebCLUST," in *IEEE Access*, pp. 1-16, 2021.
[4] P. Casas et al., "Are you on Mobile or Desktop? On the Impact of End-user Device on Web QoE Inference from Encrypted Traffic," in *CNSM*, 2020.
[5] P. Casas et al., "Mobile Web and App QoE Monitoring for ISPs - from Encrypted Traffic to Speed Index through Machine Learning," in *WMNC*, 2021.
[6] A. Huet et al., "Revealing QoE of Web Users from Encrypted Network Traffic," in *IFIP Networking*, 2020.
[7] T. Hoßfeld et al., "Speed Index: Relating the Industrial Standard for User Perceived Web Performance to Web QoE," in *QoMEX*, 2018.
[8] Q. Gao et al., "Perceived Performance of Top Retail Webpages in the Wild: Insights from Large-scale Crowdsourcing of Above-the-fold QoE," in *Internet-QoE*, 2017.
[9] D. N. da Hora et al., "Narrowing the Gap between QoS Metrics and Web QoE using Above-the-Fold Metrics," in *PAM*, 2018.
[10] F. Salutari et al., "Implications of the Multi-Modality of User Perceived Page Load Time," in *MedComNet*, 2020.
[11] R. Netravali et al., "Vesper: Measuring Time-to-Interactivity for Web Pages," in *Symposium on Networked Systems Design and Implementation (NSDI)*, 2018.
[12] S. M. Lundberg et al., "A Unified Approach to Interpreting Model Predictions," in *Int. Conference on Neural Information Processing Systems (NIPS)*, 2017.
[13] J. Schneider, "Personalization of Deep Learning," in *International Data Science Conference*, 2020.
[14] P. Meenan, "WebPageTest - Website Performance and Optimization Test," 2020. [Online]. Available: https://www.webpagetest.org/
[15] F. Iglesias et al., "Absolute Cluster Validity," in *IEEE TPAMI*, vol. 42 (9), 2020.