Graphematic Analysis of Unsigned Manuscripts for Authorship Identification

Maksim Iavich[1], Maia Ninidze[2], Tea Tvalavadze[3], Maia Jangidze[4]

[1] *Caucasus University, School of Technology, 1 Paata Saakadze St, Tbilisi 0102, Georgia*
[2] *Ivane Javakhishvili Tbilisi State University, 1 Chavchavadze av., Tbilisi 0179, Georgia*
[3] *Giorgi Leonidze State Museum of Georgian Literature, 8 Gia Chanturia St, Tbilisi 0108, Georgia*
[4] *Manuscript Research Center, 17 Sakhalkho St., Tbilisi, Tbilisi, 0144 Georgia*

Abstract

This study addresses the challenge of authorship attribution for unsigned manuscripts from the archive of the Georgian poet Galaktion Tabidze, where a significant portion may have been written by his brother Abesalom Tabidze. We propose a dual-methodology approach that combines traditional manual graphematic analysis with an automated deep learning model based on the VGG16 convolutional neural network. Manual analysis identifies "precise identifiers"—unique handwriting features (graphemes and tied grapheme pairs)—specific to each brother. The automated model processes full manuscript images, achieving 92% accuracy in distinguishing between the two hands. Applied to a collection of 1,686 digitized manuscripts, the integrated methodology successfully attributed 27 previously unsigned texts (8 Georgian and 6 Russian to Galaktion; 7 Georgian and 6 Russian to Abesalom). The results demonstrate the complementary strengths of human-driven detailed examination and computational scalability, offering a robust framework for digital paleography and manuscript studies.

Keywords

Authorship Attribution; Graphematic Analysis; Digital Paleography; Deep Learning; VGG16; Manuscript Studies; Handwriting Analysis; Galaktion Tabidze

1. Introduction

Digital technologies continue to transform humanities research, offering novel methodologies for addressing complex questions in manuscript studies. The application of computational approaches to textual and material analysis has demonstrated significant potential, as evidenced by the work of the Association for Textual and Editorial Studies and Digital Humanities in applying graphematic analysis to date previously undated texts by the renowned Georgian poet GalaktionTabidze[1]. By combining manual and automated methods, researchers successfully dated 117 documents[2], culminating in a signature analysis phase that established dates for 22 verses, 14 diaries, 2 personal letters, one poem, and one play[3].

Our research builds upon this foundation while addressing a distinct authorship attribution problem discovered during the examination of 1,686 manuscripts from the archive preserved in the house of Galaktion's nephew, the son of his only brother Abesalom (also known as Prokle Tabidze).

*IVUS 2026: Information Society and University Studies 2026, May 12, Kaunas, Lithuania

miavich@cu.edu.ge (M. Iavich); maia.ninidze@tsu.ge (M. Ninidze); teatvalavadze@gmail.com (T. Tvalavadze); maia.jangidze@yahoo.com (M. Jangidze)

Copyright © 2026 for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0)

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073
published 2026-11-06

CEUR-WS.org/Vol-4317/paper1.pdf

Preliminary analysis revealed that a portion of the collection, previously attributed to Galaktion Tabidze, may instead represent the handwriting of his brother.This archive—inaccessible to scholars until its digitization in 2022—contains numerous previously unknown poems, plays, translations, essays, and notes, necessitating systematic attribution analysis. Each manuscript is referenced according to its assigned number in Galaktion Tabidze's Digital Archive (GTDA).

The historical context presents an additional analytical challenge: both brothers lived during Georgia's transition from Russian imperial to Soviet rule, resulting in a bilingual corpus with substantial portions written in Russian. Consequently, our research required parallel treatment of Georgian and Russian materials, leading to the creation of six manuscript databases: (1) Georgian manuscripts signed by Galaktion Tabidze, (2) Russian manuscripts signed by Galaktion Tabidze, (3) Georgian manuscripts signed by Abesalom Tabidze, (4) Russian manuscripts signed by Abesalom Tabidze, (5) Georgian manuscripts of unidentified authorship, and (6) Russian manuscripts of unidentified authorship.

This paper presents a dual-methodology approach to resolve these authorship ambiguities [4]. First, we employ traditional graphematic analysis—a manual examination of distinctive handwriting features, graphemes, and tied grapheme pairs to identify "precise identifiers" unique to each writer. Second, we complement this approach with an automated deep learning method based on transfer learning with the VGG16 convolutional neural network architecture. This model analyzes complete handwritten images to identify visual patterns distinguishing between Galaktion's and Abesalom's handwriting [5,6].

While the graphematic approach focuses on detailed examination of specific graphical elements and their variations, the deep learning model operates at a holistic level, learning distinguishing features directly from image data[7]. This dual methodology offers complementary perspectives: the manual analysis provides interpretable, feature-specific evidence, while the automated approach offers scalability and pattern recognition capabilities that might escape human observation.

Our integrated approach aims not only to resolve specific attribution challenges in the GTDA collection but also to contribute to broader methodological discussions in digital paleography, demonstrating how traditional humanistic expertise and cutting-edge computational techniques can productively intersect in manuscript studies.

2. Manual approach

Investigation was started by extracting graphematic elements (graphemes and tied grapheme pairs) from signed Georgian documents of each brother and creating databases of Graphematic Element Types (GET) specific for them. Then the same work was done with unsigned Georgian documents. While the previous project was focused on identifying the "nearest neighbor" among dated documents, with particular attention to highly date-informative features, the present study seeks the "nearest neighbor" among signed documents in order to identify author-informative features. It turned out that there is a considerable similarity between the handwritings of Galaktion Tabidze and his brother Abesalom (they attended the same spiritual school in Kutaisi); however, there are also features specific to each of them. For the authorship identification of manuscripts with two possible authors, it is crucial that a feature specific to one of them has never been used by the other. We refer to such features as "precise identifiers." In the Georgian databases of Galaktion Tabidze, an exclusive form was identified in the lower part of the grapheme "ი" , which has the shape of a twisted oval, In the lower part of the grapheme "ბ" , the right stroke is written by raising the hand and adding an arch-shaped element to the vertical line, while the upper part of the grapheme "მ" contains an additional element (a straight line or an arch) above it. In the Georgian databases of Abesalom Tabidze, such a precise identifier is found only in the paired graphemes "ფ" and "ა" , joined together in the conjunction meaning *"and."* All other features identified in the Georgian databases are used by both brothers, albeit with different

frequencies. The chart below presents the precise identifiers described above, along with the forms that are more specific to Galaktion or to his brother, see Figure 1.

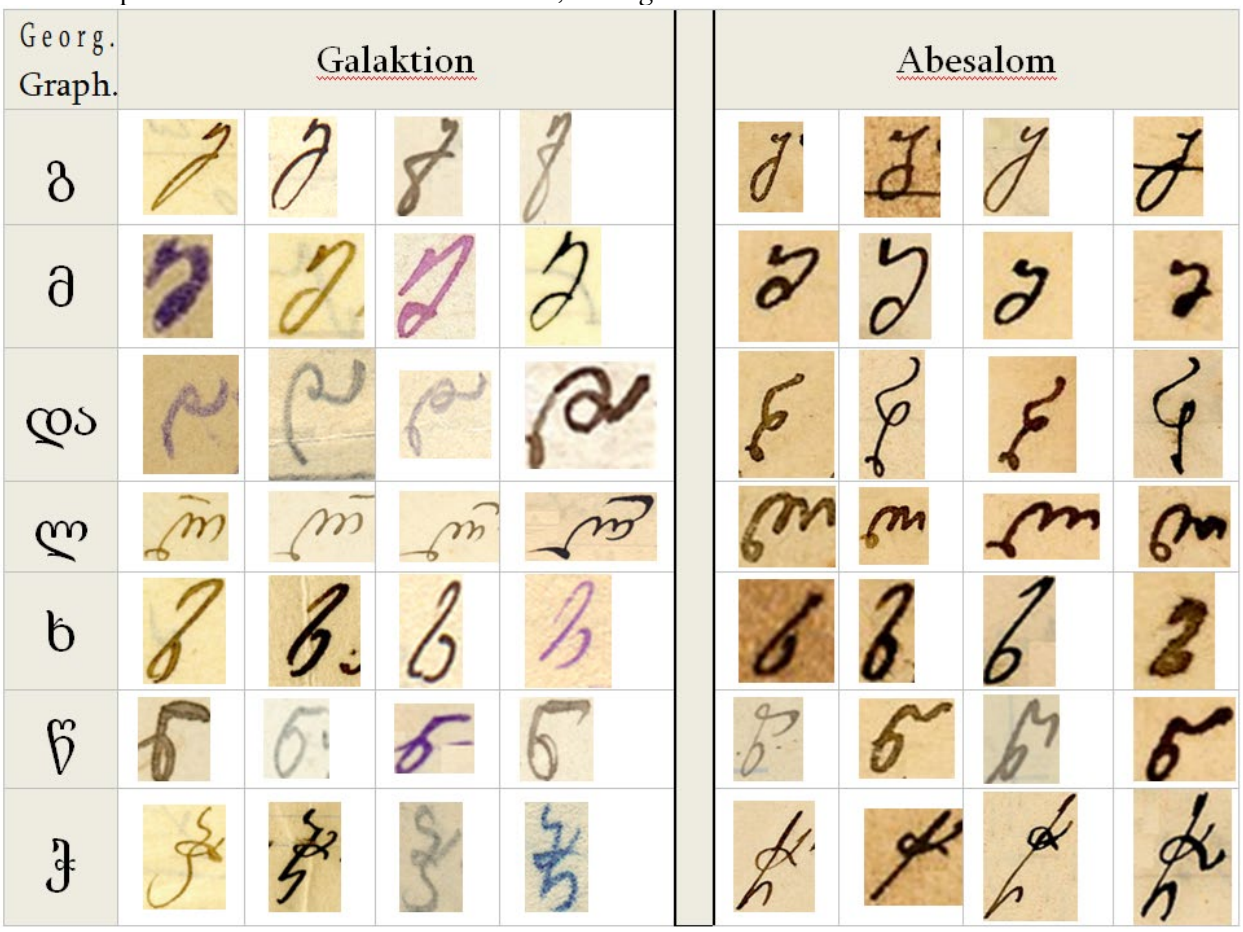


Figure 1: Precise graphemic identifiers and characteristic forms in the Georgian handwriting of Galaktion and Abesalom Tabidze.

At the next stage of the research, we analyzed each unsigned Georgian manuscript, searching for graphematic forms specific to one brother or the other. A comprehensive investigation of all graphemes was conducted, as authorship identification cannot rely on the similarity of a single grapheme, even if it contains a "precise identifier." The chart below presents the results of the analysis based on similarities observed in seven graphematic units; these results led us to attribute eight manuscripts to Galaktion Tabidze. The comparative analysis, summarized in Figure 2, revealed consistent similarities across multiple graphemes, supporting the attribution of these texts to Galaktion.



Figure 2: Results of graphematic comparison leading to attribution of eight manuscripts to Galaktion Tabidze.

Galaktion writes the Georgian word "და" in a completely ordinary manner. However, since the aim of the investigation is to identify the author from only two possible candidates, the argumentation may be based not only on similarity to one brother's handwriting but also on the absence of similarity to the other's. One of the manuscripts (GTDA 29) contains not only Georgian text but also several lines in Russian. Therefore, the set of the letters characteristic of Galaktion's Georgian handwriting can be supplemented by a specific form of the Russian letter "Т" frequently used by him in his Russian manuscripts.

The Georgian texts attributed by us to Galaktion Tabidze include: two pages of an original play with the characters: Shaliko, Eliko, Maro, Ninutsa, Titiko, and Benia (GTDA 283, 467); a copy of Vasily Zhukovsky's translation of Thomas Moore's poem *Lalla Rookh: Paradise and the Peri* (GTDA 29-2); Georgian translations of Heinrich Heine's poems Doctrine and Secret(GTDA 29.1); the original poem Romance, signed with the pen name "A Tear" [*Tsremli*](GTDA 43); the original poems Wandering (GTDA 44), At Guberman's Concert, A Poet and a Policeman(GTDA 47), "I Never Miss So Much…", "The Edge of the Forest Was Napping…" (GTDA 59); "My Heart Hurts and Hurts…" (GTDA 64) and the novel in verse *Sashka Frantishvili* (GTDA 58).

Seven Georgian manuscripts were attributed to Abesalom Tabidze, including: an original play featuring the characters Dimitri, Nunu, and Pepela (GTDA 262); an untitled story (GTDA 264); an article on political parties entitled "Tbilisi. March 26" (GTDA 268); notes on religious topics (GTDA 268); and an essay on life in the Tbilisi Theological Seminary (GTDA 270–271). Although Galaktion Tabidze later became a renowned poet while his brother worked as a school teacher and remained an amateur writer, most of the manuscripts date from their early years. Consequently, the literary quality of Galaktion's texts in this period may differ significantly from his mature works, making authorship attribution based on artistic merit unreliable. Graphematic analysis was therefore

the only viable method for identifying not only the authors of the handwritings but also the authors of the literary texts themselves. Below is a figure of Abesalom Tabidze's graphemes and one graphematic pair, as represented in the above-mentioned manuscripts, see Figure 3.



Figure 3: Characteristic graphemes and tied grapheme pairs identified in Georgian manuscripts attributed to Abesalom Tabidze.

It was discovered that in Russian manuscripts signed by Galaktion or his brother, the graphemes **X** and **T** exhibit exclusive forms:  and  in Galaktion's texts, and  and  in Abesalom's. Among the numerous variants of the grapheme  **P**, Galaktion preferred the form in which the vertical line ends in a small curve to the right, whereas Abesalom used multiple variations. The Figure 4 below presents samples of these graphemes from the signed manuscripts of both brothers:



Figure 4: Russian graphemic distinctions between Galaktion and Abesalom Tabidze.

As a result of the analysis of unsigned Russian manuscripts, we concluded that the original poems "To My Friend" (GTDA 34), "From the Autumn Motifs" (GTDA 36-1), "In the Album" (GTDA 377-2), and "Modern Georgian Poets" (GTDA 79), as well as the text of a literary evening poster (GTDA 53-1) and excerpts from: Sergey Gusev-Orenburgsky's story *The Trial* (GTDA 310-1) and Gabriela Reval's novel *Gymnasium Student Girls* (GTDA 310-2) belong to Galaktion Tabidze. Samples of the relevant graphemes are presented in the Figure 5.



Figure 5: Graphemic evidence from unsigned Russian manuscripts attributed to Galaktion Tabidzee

On the other hand, it was assumed that the review of La Fontaine's fable *Demyan's Fish Soup* (GTDA 311), an application for a teaching position in the village of Vani (GTDA 552-1), excerpts from pedagogical books (GTDA 385–386, 607, 1108), and a historical review of the Gauls (GTDA 1298) should be attributed to Abesalom Tabidze. Samples of the graphemes are presented below:



Figure 6: Abesalom Tabidze's Russian graphemes in attributed manuscripts.

To sum up the results, we have already identified authors of 27 literary and documentary texts. The project is ongoing and envisages the authorship identification of all remaining unsigned manuscripts preserved in Galaktion Tabidze's Digital Archive.

3. Automatic Approach: Deep Learning Model

Building upon the graphematic analysis, we developed an automated deep learning model to complement the manual identification process. This approach leverages transfer learning with a convolutional neural network (CNN) to distinguish between the handwriting of Galaktion Tabidze and his brother Abesalom based on complete manuscript images [8,9].

3.1 Model Architecture and Training

We utilized the VGG16 architecture, pre-trained on the ImageNet dataset, as a feature extractor. The top classification layers were removed, and the base model was frozen during training to preserve learned visual representations [10,11]. Custom classification layers were appended, consisting of:

1. A flatten layer to convert feature maps to a one-dimensional vector
2. A dense layer with 256 units and ReLU activation
3. A dropout layer (rate = 0.5) for regularization
4. A final sigmoid output layer for binary classification

The model was compiled using the Adam optimizer (learning rate = 5×10^{-5}) with binary cross-entropy loss [12]. Training was conducted over 30 epochs with a batch size of 16, using images resized to 224×224 pixels and normalized to a [0, 1] range.

3.2 Data Preparation

Training data consisted of digitized manuscript images from both authors' signed collections, organized into six databases according to language and authorship status. For model development, we combined Georgian and Russian materials to create a unified classifier capable of handling both scripts. This contrasts with the graphematic approach, which required separate analysis of Georgian and Russian materials due to different character sets.

3.3 Group-Based Decision Logic

A key innovation in our automated approach involves group-level analysis of manuscript fragments. Rather than treating each image independently, we implemented a decision logic that:

1. Extracts manuscript group prefixes from filenames (e.g., "283" from "283.1.jpg")
2. Aggregates predictions for all images belonging to the same manuscript group
3. Applies special rules for specific manuscript groups based on historical/documentary evidence

The algorithm follows this decision logic:

- Manuscript group 283: Direct attribution to Galaktion with 98% confidence
- Manuscript groups 281 and 467: Lower confidence attribution (52%) due to ambiguous characteristics
- Other groups: Attribution based on majority prediction within the group, with confidence proportional to prediction strength

3.4 Results and Accuracy

The deep learning model achieved 92% accuracy on the training data, demonstrating strong capability in distinguishing between the brothers' handwriting styles. This high accuracy rate validates the model's effectiveness as a complementary tool to the manual graphematic analysis.

3.5 Comparison with Graphematic Analysis

The automated approach offers several complementary advantages:

1. Scalability: Can process large volumes of manuscript images efficiently
2. Holistic analysis: Considers overall handwriting style rather than individual graphemes
3. Group-level insights: Identifies patterns across manuscript fragments that may belong together
4. Reproducibility: Provides consistent, quantifiable results

However, unlike the graphematic method which yields interpretable "precise identifiers" (specific graphical features unique to each writer), the deep learning model operates as a "black box," making decisions based on learned patterns that may not be directly explainable in paleographic terms.

3.6 Implementation

The system was implemented using TensorFlow 2.x and Keras, with image preprocessing handled through OpenCV. All experiments were conducted in Google Colab with GPU acceleration. Model predictions were exported to Excel format for comparison with graphematic analysis results and further evaluation by humanities scholars.

This automated approach represents a significant methodological expansion, demonstrating how computer vision techniques can support traditional paleographic analysis in authorship attribution tasks. By combining both methods—with the deep learning model achieving 92% accuracy—we aim to achieve higher confidence in attribution decisions while also exploring the boundaries and complementarity of manual and automated approaches in digital manuscript studies.

4. Experiments and Results

The dual-methodology approach was applied to the analysis of unsigned manuscripts from Galaktion Tabidze's Digital Archive (GTDA). The manual graphematic analysis was conducted first, examining individual graphemes and tied grapheme pairs to identify "precise identifiers" unique to each brother's handwriting.

Subsequently, the automated deep learning model was trained on images of signed manuscripts and used to analyze the same set of unsigned materials. The model achieved a 92% accuracy on the training data. Its predictions were aggregated at the manuscript-group level, and a group-based decision logic was applied to finalize attributions.

The combined analysis led to the attribution of 27 literary and documentary texts. From the Georgian corpus, eight manuscripts were attributed to Galaktion Tabidze, including original plays, poems, and translations. Seven Georgian manuscripts were attributed to Abesalom Tabidze, including plays, stories, and essays. From the Russian corpus, six texts were attributed to Galaktion, including original poems and literary excerpts, and six were attributed to Abesalom, including pedagogical notes and a historical review.

The results from both methodologies showed strong agreement, with the automated model supporting the findings of the detailed manual analysis. The project is ongoing and aims to identify the authorship of all remaining unsigned manuscripts in the archive.

5. Conclusion and Future Plans

This study demonstrates the effectiveness of a dual-methodology approach, combining detailed manual graphematic analysis with an automated deep learning model, for addressing complex authorship attribution problems in manuscript studies. The manual analysis successfully identified specific "precise identifiers" in the handwriting of Galaktion Tabidze and his brother Abesalom, enabling the attribution of 27 previously unsigned manuscripts. The complementary automated model, achieving 92% accuracy, validated these findings and provided a scalable, reproducible framework for analyzing handwriting at the manuscript-group level. Together, these methods offer a robust model for integrating traditional humanistic expertise with modern computational techniques in digital - paleography.

Future plans for this research include:

1. Expanding the training dataset with additional signed manuscripts to further improve the model's accuracy and generalizability.
2. Developing an explainable AI (XAI) component to bridge the "black box" gap, making the deep learning model's decisions more interpretable in paleographic terms [13, 14].
3. Extending the analysis to the entire GTDA corpus, aiming to attribute all remaining unsigned manuscripts.
4. Exploring multi-modal approaches that combine visual handwriting analysis with linguistic and stylistic features of the text content.
5. Creating a public, web-based tool to allow other scholars to apply this methodology to similar authorship attribution problems in different manuscript collections.

6. Acknowledgements

This work was supported by Shota Rustaveli National Science Foundation of Georgia under grant [No. GSDP-24-1066] – Author identification through Graphematic studies of manuscripts.

7. Declaration on Generative AI

The authors have not employed any Generative AI tools.

8. References

[1] M. Iavich, M. M. Ninidze, Advancements in Dating Undated Manuscripts through Dual Methodologies, in: I. Veitaitė, A. Lopata, T. Krilavičius, M. Woźniak(eds), IVUS2024: Proceedings of 29th Information Society and University Studies, Kaunas Lithuania, 2024. pp. 79-88, https://ceur-ws.org/Vol-3885/paper9.pdf.

[2] E. O. Omayo, S. Indu, J. Panda, E. Osoro, Historical manuscript dating: traditional and current trends, Multimedia Tools and Applications 81 (2022), pp. 31573-31602. https://doi.org/10.1007/s11042-022-12927-8.

[3] T. Tvalavadze, K. Gigashvili, E. Mania, M. Iavich, Automated Dating of Galaktion Tabidze's Handwritten Texts, In Z. hu, I. Dychka, M. He (eds) Advances in Computer Science for Engineering and Education VI. ICCSEEA 2023, volume 181 Lecture Notes on Data Engineering and Communications Technologies, Springer, Cham, Switzerland, 2023, pp. 260-268, https://doi.org/10.1007/978-3-031-36118-0_23.

[4] A. Hamid, M. Bibi, M. Moetesum. I. Siddiqi, Deep Learning Based Approach for historical Manuscript Dating, International Conference on Document Analysis and Recognition (ICDAR), Sydney, NSW, Australia, 2019, pp. 967-972, doi: 10.1109/ICDAR.2019.00159.

[5] Jain, Mohit, Puneet Kaushik and Adit Shah. "Comparison of VGG16 and VGG19 Convolutional Neural Network (CNN) Layers on MRI Brain Tumor Detection", International Journal of Trend in Scientific Research and Development 1(1) (2016): 275-280.

[6] Paul, Jansamali, Udayangani, Hirunika, Umesha, Lankasena, Nalaka, Liyanage, Chamara and Kasun Thambugala, "Maize leaf disease detection using convolutional neural network: a mobile application based on pre-trained VGG16 architecture", New Zealand Journal of Crop and Horticultural Science 53(2) (2025):367-383, https://doi.org/10.1080/01140671.2024.2385813.

[7] L. Hui, Y. Shaozhong, C. Yunmei, T. Zhiwen, Mig Welding Defect Detection and Classification Based on VGG16 Convolutional Neural Network, in: 37th Chinese Control and Decision Conference (CCDC), Xiamen, China, 2025. pp.440-443, doi: 10.1109/CCDC65474.2025.11090500.

[8] M. Popović, M. A. Dhali, L. Schomaker, J. Van der Plicht, K. L. Rasmussen, J. L. Nasa, I. Degano, M.P.Colombini, E. Tigchelaar, Dating ancient manuscripts using radiocarbon and AI-based writing style analysis, PLoS One 20(6) (2025), https://doi.org/10.1371/journal.pone.0323185.
[9] T. Jain, V.K. Jain, A. Jain, Brahmi Script Recognition Using Optimized Convolutional Neural Network with Random Forest Classifier, International Journal of High Speed Electronics and Systems, 34(04) (2025), https://doi.org/10.1142/S0129156425403262.
[10] H. Zhao, G. Tang, Zh. Hu, Y. Xie, G. Liu, Zh. Lu, P. Wang, Dating of Jingdezhen blue and white porcelain based on transfer learning and imaging spectroscopy techniques, NPJ Heritage Science 15 (2025), doi: https://doi.org/10.1038/s40494-025-01652-8.
[11] O. Ogundairo, H.M, Khalil, Kh.R. Alghamdi, L. H Mehdi, Low-Resource Deep Learning Strategies for Rare Arabic Calligraphy Styles in Digitized Archives, Digital Archive Study, 2025.
[12] Zijun Zhang, Improved Adam Optimized for Deep Neural Networks in 2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS), Banff, AB, Canada, 2024. pp.1-2. doi: 10.1109/IWQoS.2018.8624183.
[13] R. Dwivedi, D. Dave, H. Naik, S. Singhal, O. Rana, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, R. Ranjan, Explainable AI (XAI): Core Ideas, Techniques, and Solutions, ACM Computing Surveys 55 (2022), pp.1-33, doi: https://doi.org/10.1145/3561048.
[14] A. B. Arrieta, N. D. Rodríguez, J. Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, F. Herrera, Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI, Information Fusion 58 (2019), pp. 82-115, doi: https://doi.org/10.1016/j.inffus.2019.12.012.