iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 13 **Issue:** III **Month of publication:** March 2025

DOI: https://doi.org/10.22214/ijraset.2025.67269

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 13 Issue III Mar 2025- Available at www.ijraset.com

Vershachi Unlearning: A Framework for Machine Unlearning

Mr. Veerasagar S S
Graduate student, Department of Computer Science and Engineering, Govt. SKSJT Institute, KR Circle, Bangalore, Karnataka, India

***Abstract:** In the contemporary landscape of the digital world where industry relies on the technology of artificial intelligence which fundamentally depends on the concepts of machine learning. Machine learning is a field where it utilizes the immense amount of data and then feeds this data into a structure called models. This data "trains" this model. Abundant data is used to train these models, for this data to be as accurate as it can be optimally. However, reliance on this abundant data exposes us to a significant risk to user privacy which is a matter of concern. It directly challenges the existence of "right to be forgotten". There is an intricate relation between the model and the data with which it is trained. Traditional data management systems can easily erase user information from databases, but the scenario becomes considerably complex with machine learning models. This gives rise to the whole new concept called machine unlearning. This project addresses this challenge by developing a standalone tool and API specifically designed to facilitate the forgetting of data by machine learning models. Our objective is to pioneer a practical approach to enhance user privacy in the context of machine learning technologies. By creating an efficient and reliable solution, we aim to bridge the gap between data privacy rights and the intricate workings of machine learning models. Through this endeavor, we contribute to the evolving discourse on privacy, data security, and ethical AI practices in the digital age.*

I. INTRODUCTION

In current trends, in abundance the data generated by users is being accumulated. There is a risk in the data being compromised without the consent of the users. The recent regulations now require that, on request, private information about a user must be removed from both computer systems and from ML models. These are the results of new privacy protection laws like General Data Protection Regulation (GDPR), California Consumer Privacy Act (CCPA) and many more that are governing the privacy of the users. These laws introduce a concept called "right to be forgotten." wherein the data of the users must be deleted from the databases and ML models when the user wishes it. To comply with these guidelines the company or the developer must be in a place to remove the data from the server and databases. To remove the data from the back-end databases is relatively easier than removing the data from the machine learning models. We can try retraining the models from scratch, whereas in this case the training is done excluding all the users' personal data. But it is too expensive both timewise and computational wise. There is another scenario where one can selectively remove the data from the ML models while still preserving the original accuracy. This introduces a new domain called "Machine Unlearning". Our project aims to make a framework for machine unlearning.

A. *Existing Systems*

In the existing system analysis, the work done is:

1) They require access to the original data or a proxy of it, which may not be available or feasible in some scenarios
2) They incur high computational and storage costs, especially for complex models and large datasets.
3) They may not guarantee complete forgetting or privacy preservation, as some traces of the unlearned data may remain in the model or its outputs.
4) They may affect the performance and accuracy of the model, as unlearning some data may reduce the generalization ability of the model or introduce biases.

B. *Proposed System*

This is the system we propose:

1) The proposed framework for machine unlearning is a novel approach that allows users to remove unwanted data from trained models without retraining them from scratch.
2) The framework is developed in python and supports popular machine learning libraries such as TensorFlow, PyTorch, and scikit-learn. It also provides a user-friendly interface and documentation to help users get started with machine unlearning.

3) The framework is based on the theoretical foundations of machine unlearning and guarantees that the unlearned model will have similar performance and generalization as the original model, while satisfying the privacy and fairness constraints of the users.

4) This framework is a valuable tool for machine learning practitioners and researchers who want to incorporate machine unlearning into their workflows and applications. It can help them address the challenges of data quality, data ownership and data regulation

II. SCOPE

A Machine Unlearning project aims to develop techniques and methodologies for a machine learning model to selectively forget or update previously learned information. The objectives and scope of such a project typically include Selective Forgetting, Ethical Consideration, Security, Privacy and User Interaction.

III. DESIGN

A. System Design

Finally, the system design process often entails iterative refinement based on feedback, testing, and evolving requirements, ensuring the system meets its objectives effectively and adaptively.

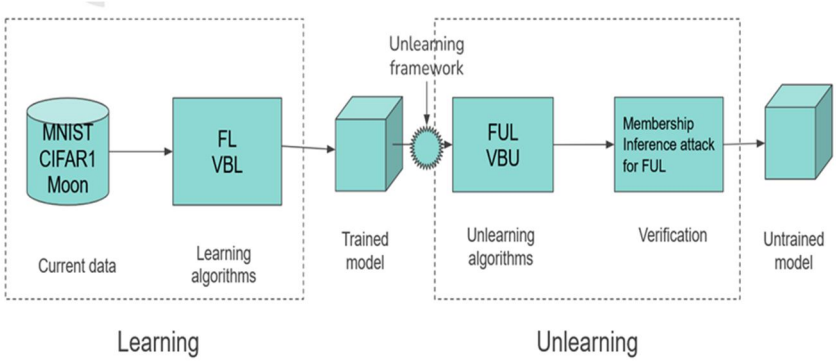


Fig 3.1: The process of unlearning the trained data

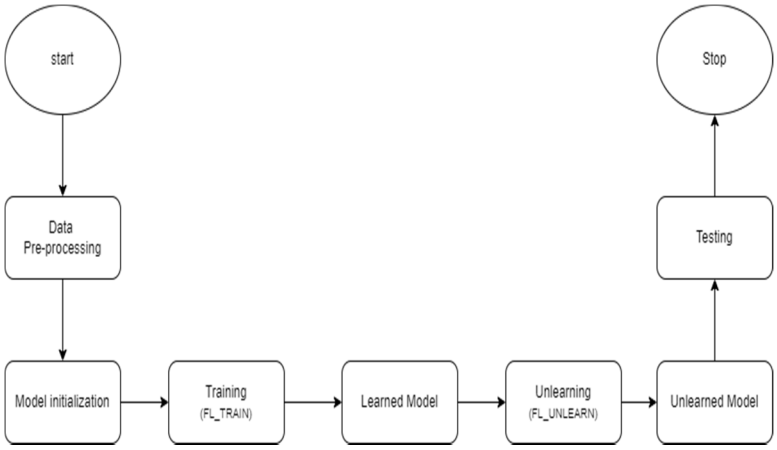


Fig 3.2: Flow diagram of machine unlearning

Designing a system for learning involves steps like data ingestion, preprocessing, model training, evaluation, and deployment. Data is collected and processed, features are engineered, and models are trained using various algorithms. Model performance is assessed using metrics like accuracy, and once satisfactory, the model is deployed for predictions. In contrast, unlearning involves identifying and modifying previously learned information in the model parameters. Techniques like parameter adjustment or removal are applied, followed by performance evaluation and potential redeployment, with considerations for transparency and regulatory compliance.

B. Federated Unlearning

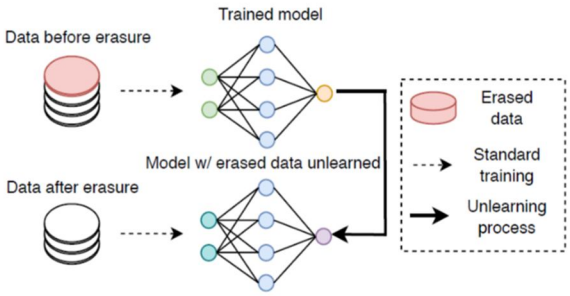


Fig 3.3: Federated Unlearning

In federated learning, the primary objective is indeed to train a global model while keeping the raw data decentralized and private on client devices. This way, data remains on the devices, reducing privacy concerns associated with centralized data collection.

Now, let us delve into the concept of unlearning and its specific parameter, unlearn_global_model.

1) Unlearning in Federated Learning: Unlearning in federated learning refers to the process of removing a client's contribution from the global model. This is important for various reasons, such as when a client wants to stop participating in the federated learning process or if their data becomes outdated or irrelevant.

2) Unlearning Global Model Parameter: The unlearn_global_model parameter likely refers to a mechanism for initiating the unlearning process across the entire federated learning system, specifically targeting the global model. Here is a breakdown:

- Unlearn Local Model vs. Global Model: Before understanding unlearn global model, it is essential to differentiate between unlearning a local
  model and unlearning the global model. When a client unlearns its local model, it removes its specific contributions to the global model. However, unlearning the global model entails removing the influence of a client from the collective global model across all participating clients.
- Purpose of Unlearning Global Model: The unlearn global model parameter allows for a coordinated process where a client requests the removal of its data from the global model across the entire federate learning system. This could be crucial for compliance with privacy regulations or in situations where clients want to ensure that their data no longer contributes to the global model.
- Mechanism for Data Removal: Implementing unlearn global model would likely involve communication between the client and the server. When a client triggers this parameter, it signals to the server that its data should be disregarded in the next global model update. The server then orchestrates the removal of this client's data from the global model during the aggregation process.
- Impact on Model Performance: Unlearning a client's contribution from the global model might impact on the overall performance of the model, especially if that client had provided significant and relevant data. Therefore, mechanisms for unlearning should be carefully managed to minimize disruptions to model performance while upholding privacy and data integrity.

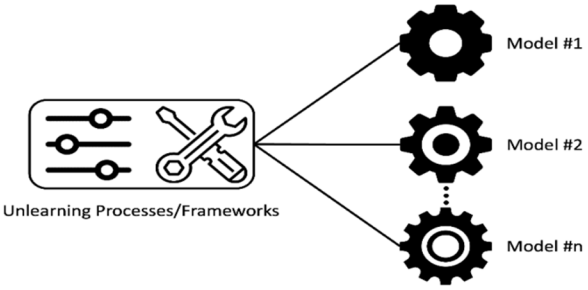
3) Federated Averaging (FedAvg): is a key algorithm in federated learning where multiple decentralized clients collaboratively train a global model while keeping their data private. In the context of machine unlearning, FedAvg plays a significant role in enabling the removal of individual client contributions from the global model. When a client wishes to unlearn or remove its data from the global model due to privacy concerns or other reasons, FedAvg allows for this by updating the global model based on the remaining clients' contributions, effectively mitigating the impact of the departing client's data. By averaging the model updates from the remaining clients, FedAvg ensures that the global model continues to evolve without retaining the specific information contributed by the departing client.

When a client initiates the unlearning process, FedAvg facilitates the removal of its data from the global model by updating the model based on the contributions of the remaining clients.

IV. IMPLEMENTATION

Types of unlearning algorithms:

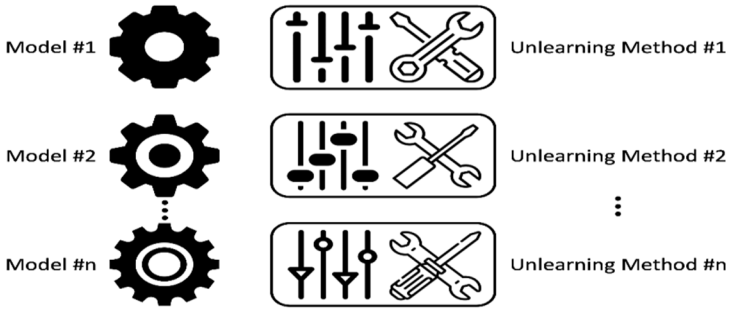
1) Model-Agnostic Unlearning



Model-agnostic unlearning algorithms are designed to be versatile and adaptable across various machine learning models. They operate independently of the internal structure or architecture of the model, focusing solely on the training data. Here's a deeper look into how these algorithms work:

- Algorithmic Approach: Model-agnostic unlearning algorithms typically employ techniques that analyze the training data to identify specific data points for removal. These algorithms prioritize the removal of data points that are deemed irrelevant, outdated, or potentially sensitive.
- Data-Driven Analysis: Instead of relying on the intricacies of a particular model, model-agnostic algorithms leverage statistical methods and data mining techniques to analyze patterns within the training data. This data-driven approach allows for the identification of data points that contribute minimally to the model's predictive performance or may pose privacy risks.
- Flexibility and Compatibility: One of the key advantages of model-agnostic unlearning is its flexibility and compatibility with different machine learning models. Whether the model is based on deep learning, decision trees, support vector machines, or any other algorithm, model-agnostic algorithms can be applied seamlessly without requiring modifications to the model's architecture.

2) Model Intrinsic Unlearning



Model intrinsic unlearning algorithms are specifically tailored to the internal workings and architecture of a particular machine learning model. These algorithms leverage a deep understanding of the model's structure to achieve precise and optimized unlearning outcomes. Here's a deeper exploration of how model intrinsic unlearning works:

- Model-Specific Optimization: Unlike model-agnostic approaches, model intrinsic unlearning algorithms are intricately tied to the underlying architecture of the model. These algorithms are designed to exploit the unique characteristics and properties of the model to achieve more efficient and effective unlearning.
- Fine-Grained Control: Model intrinsic algorithms offer fine-grained control over the unlearning process, allowing for targeted removal of specific data points or features that may impact the model's performance. By leveraging insights into the model's internal representations, these algorithms can identify and eliminate data points that contribute to biases, overfitting, or other performance issues.
- Optimization Techniques: Model intrinsic unlearning often involves the use of optimization techniques tailored to the model's architecture. These techniques may include gradient-based methods, regularization strategies, or specialized loss functions designed to minimize the impact of data removal on the model's overall performance.

Let's delve deeper into each type of unlearning algorithm to provide a more detailed understanding:

a) Model-Agnostic Unlearning

Model-agnostic unlearning algorithms are designed to be versatile and adaptable across various
machine learning models. They operate independently of the internal structure or architecture of the model, focusing solely on the training data. Here's a deeper look into how these algorithms work:

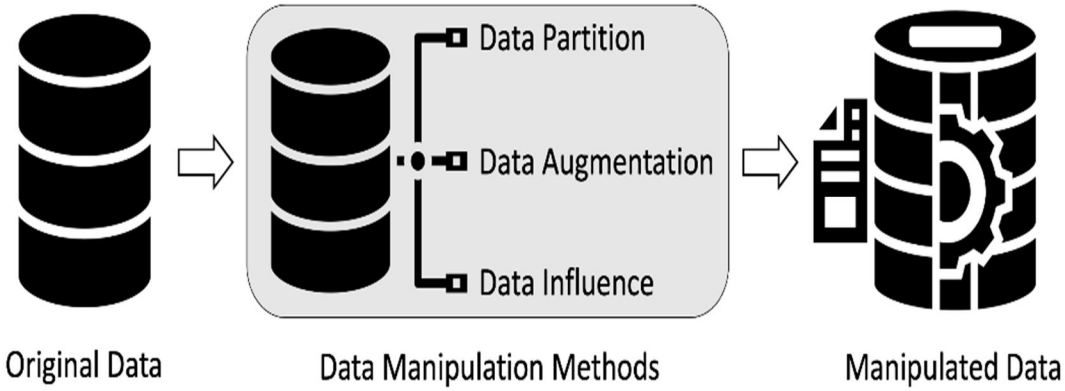
- Algorithmic Approach: Model-agnostic unlearning algorithms typically employ techniques that analyze the training data to identify specific data points for removal. These algorithms prioritize the removal of data points that are deemed irrelevant, outdated, or potentially sensitive.
- Data-Driven Analysis: Instead of relying on the intricacies of a particular model, model-agnostic algorithms leverage statistical methods and data mining techniques to analyze patterns within the training data. This data-driven approach allows for the identification of data points that contribute minimally to the model's predictive performance or may pose privacy risks.
- Flexibility and Compatibility: One of the key advantages of model-agnostic unlearning is its flexibility and compatibility with different machine learning models. Whether the model is based on deep learning, decision trees, support vector machines, or any other algorithm, model-agnostic algorithms can be applied seamlessly without requiring modifications to the model's architecture.

b) Model Intrinsic Unlearning

Model intrinsic unlearning algorithms are specifically tailored to the internal workings and architecture of a particular machine learning model. These algorithms leverage a deep understanding of the model's structure to achieve precise and optimized unlearning outcomes. Here's a deeper exploration of how model intrinsic unlearning works:

- Model-Specific Optimization: Unlike model-agnostic approaches, model intrinsic unlearning algorithms are intricately tied to the underlying architecture of the model. These algorithms are designed to exploit the unique characteristics and properties of the model to achieve more efficient and effective unlearning.
- Fine-Grained Control: Model intrinsic algorithms offer fine-grained control over the unlearning process, allowing for targeted removal of specific data points or features that may impact the model's performance. By leveraging insights into the model's internal representations, these algorithms can identify and eliminate data points that contribute.
- Optimization Techniques: Model intrinsic unlearning often involves the use of optimization techniques tailored to the model's architecture. These techniques may include gradient-based methods, regularization strategies, or specialized loss functions designed to minimize the impact of data removal on the model's overall performance.

3) Data-Driven Unlearning



Data-driven unlearning algorithms focus on analysing patterns within the training data itself to identify specific data points that need to be removed. These algorithms prioritize the removal of data points that are deemed irrelevant, outdated, or potentially sensitive.

V. RESULTS

A. Federated Unlearning

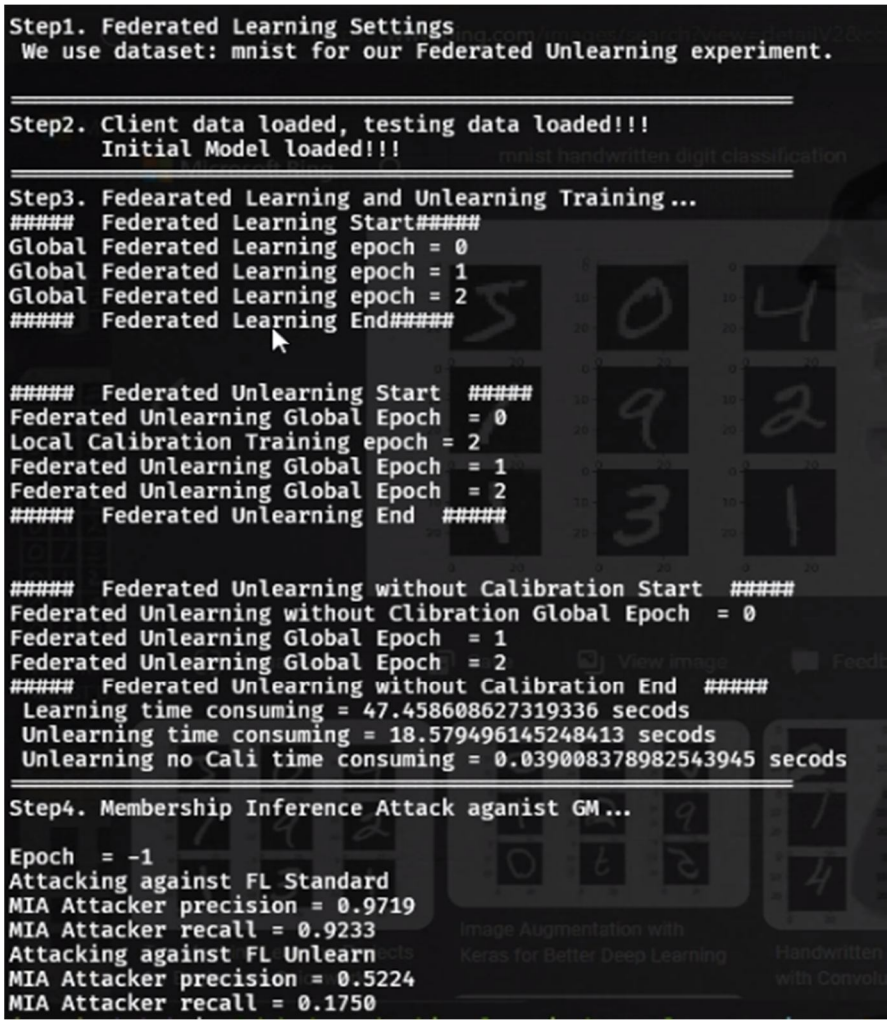


Fig 5.1: Output of federated unlearning

1) Time for Learning Process: The "time for learning process" refers to the duration required to train a machine learning model on a dataset. This encompasses tasks such as data preprocessing, model training, validation, and possibly hyperparameter tuning.
2) Time for Unlearning Process: The "time for unlearning process" refers to the duration required to remove or update specific knowledge or information from a machine learning model. The time for the unlearning process involves identifying the data to be removed, retraining the model without that data, and recalibrating model parameters.
3) Time for Unlearning without Calibration Process: "Time for unlearning without calibration process" denotes the duration needed to remove or update specific knowledge from a machine learning model without subsequent recalibration of model parameters.
4) Recall and Precision: Recall and precision are two fundamental metrics used to evaluate the performance of classification models, particularly in binary classification tasks. Recall measures the model's ability to correctly identify all relevant instances from a dataset. It calculates the ratio of true positives (correctly predicted positive instances) to the sum of true positives and false negatives (missed positive instances). Precision, on the other hand, measures the model's accuracy in predicting positive instances. It calculates the ratio of true positives to the sum of true positives and false positives (incorrectly predicted positive instances).

Time for learning process	**47.458s**
Time for unlearning process	**18.579s**
Time for unlearning without calibration process	**0.03900s**
Federated learning	**Federated unlearning**
Recall - 0.9719	**Recall - 0.5224**
Precision -0.9233	**Precision - 0.1750**

Fig 5.2: Table for the Federated Unlearning results

B. Variational Bayesian Unlearning

Variational Bayesian Unlearning is a technique used in machine learning to selectively remove or "forget" specific data points from a model trained using variational Bayesian methods. Unlike traditional unlearning methods, which often require retraining the entire model from scratch, variational Bayesian unlearning offers a more efficient approach by updating the model's parameters to accommodate the removal of specific data points. The results of variational Bayesian unlearning typically involve assessing the impact of data removal on the model's performance and updating the model's parameters accordingly.

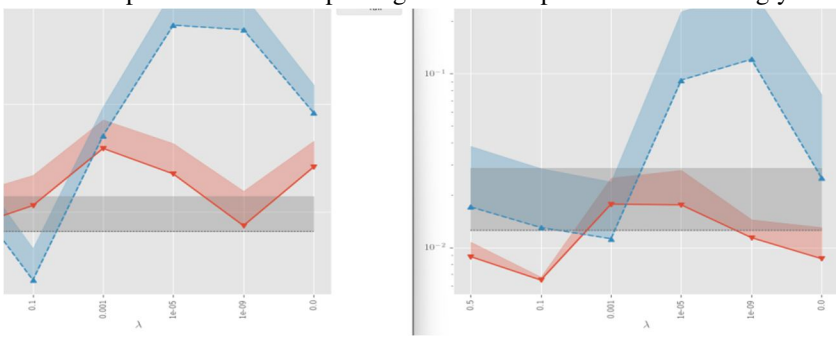


Fig 5.3: VBU graph

1) Comparison and Analysis

Comparing the pre- and post-unlearning data distributions allows us to assess the impact of unlearning on the dataset's statistical characteristics. This comparison may involve analyzing changes in the distribution's central tendency, spread, shape, or other relevant properties.

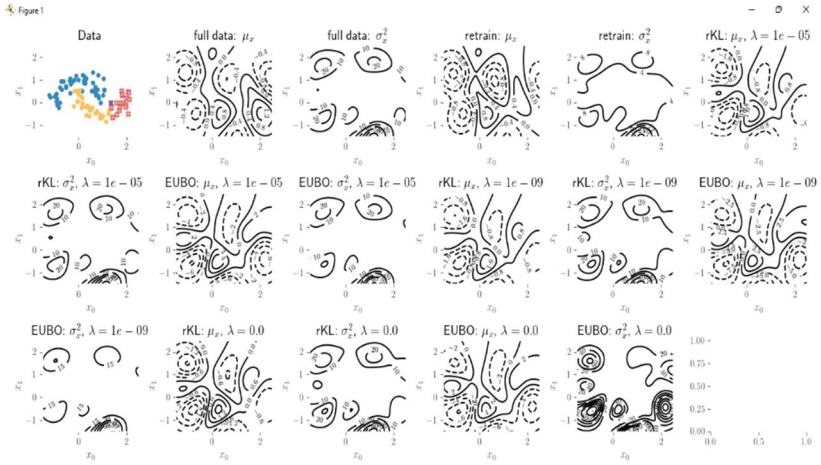


Fig 5.4: Comparison of data distribution

VI. CONCLUSION

A. *Framework Conclusion*

In summary, the machine unlearning framework represents a significant advancement in machine learning, offering efficient solutions for addressing data privacy concerns and enabling the selective removal of data from models. Its ability to provide different tools and combine various algorithms according to specific requirements makes it a versatile and powerful tool for managing and controlling data while maintaining model integrity and performance.

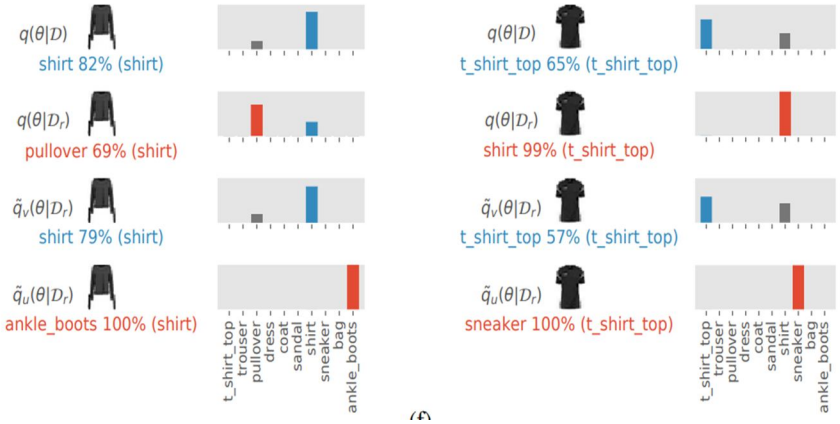


Fig 6.1: Classification results

REFERENCES

[1] A Survey Of Machine Unlearning - Nguyen, T. T., Huynh, T. T., Nguyen, P. L., Liew, A. W. C., Yin, H., & Nguyen, Q. V. H. (2022). A survey of machine unlearning. arXiv preprint arXiv:2209.02299

[2] Coded Machine Unlearning - Aldaghri, N., Mahdavifar, H., & Beirami, A. (2021). Coded machine unlearning. IEEE Access, 9, 88137-88150

[3] Toward Highly-Efficient and Accurate Services QoS Prediction via Machine Unlearning - Zeng, Y., Xu, J., Li, Y., Chen, C., Dai, Q., & Du, Z. (2023). Towards Highly-efficient and Accurate Services QoS Prediction via Machine Unlearning. IEEE Access

[4] Approximate Data Deletion from Machine Learning Models - Izzo, Z., Smart, M. A., Chaudhuri, K., & Zou, J. (2021, March). Approximate data deletion from machine learning models. In International Conference on Artificial Intelligence and Statistics (pp. 2008-2016). PMLR.

[5] Fast Yet Effective Machine Unlearning - Tarun, A. K., Chundawat, V. S., Mandal, M., & Kankanhalli, M. (2023). Fast yet effective machine unlearning. IEEE Transactions on Neural Networks and Learning Systems. arXiv:2111.08947 [cs.LG]

[6] Lifelong Anomaly Detection Through Unlearning - Min Du, Zhi Chen, Chang Liu, Rajvardhan Oak, and Dawn Song. 2019. Lifelong Anomaly Detection Through Unlearning. In 2019 ACM SIGSAC Conference on Computer and Communications Security (CCS '19), November 11–15, 2019, London, United Kingdom. ACM, New York, NY, USA, 15 pages.

[7] Machine Unlearning: Its Nature, Scope, and Importance for a "Delete Culture" - Floridi, L. (2023). Machine Unlearning: its nature, scope, and importance for a "delete culture". Philosophy & Technology, 36(2), 42.

[8] "Amnesia" - A Selection of Machine Learning Models That Can Forget User Data Very Fast - Schelter, S. (2020). Amnesia-a selection of machine learning models that can forget user data very fast. suicide, 8364(44035), 46992.

[9] Efficient Repair of Polluted Machine Learning Systems via Causal Unlearning - Yinzhi Cao, Alexander Fangxiao Yu, Andrew Aday, Eric Stahl, Jon Merwine, and Junfeng Yang. 2018. Efficient Repair of Polluted Machine Learning Systems via Causal Unlearning. In Proceedings of 2018 ACM Asia Conference on Computer and Communications Security, Incheon, Republic of Korea, June 4–8, 2018 (ASIA CCS '18), 13 pages.

[10] Remember what you want to forget - Sekhari, A., Acharya, J., Kamath, G., & Suresh, A. T. (2021). Remember what you want to forget: Algorithms for machine unlearning. Advances in Neural Information Processing Systems, 34, 18075-18086

[11] When Machine Unlearning Jeopardizes Privacy - Chen, Min, et al. "When machine unlearning jeopardizes privacy." Proceedings of the 2021 ACM SIGSAC conference on computer and communications security. 2021. arXiv:2005.02205 [cs.CR]

[12] Towards making systems forget with machine unlearning - Y. Cao and J. Yang, "Towards Making Systems Forget with Machine Unlearning," 2015 IEEE Symposium on Security and Privacy, San Jose, CA, USA, 2015, pp. 463-480, doi: 10.1109/SP.2015.35.

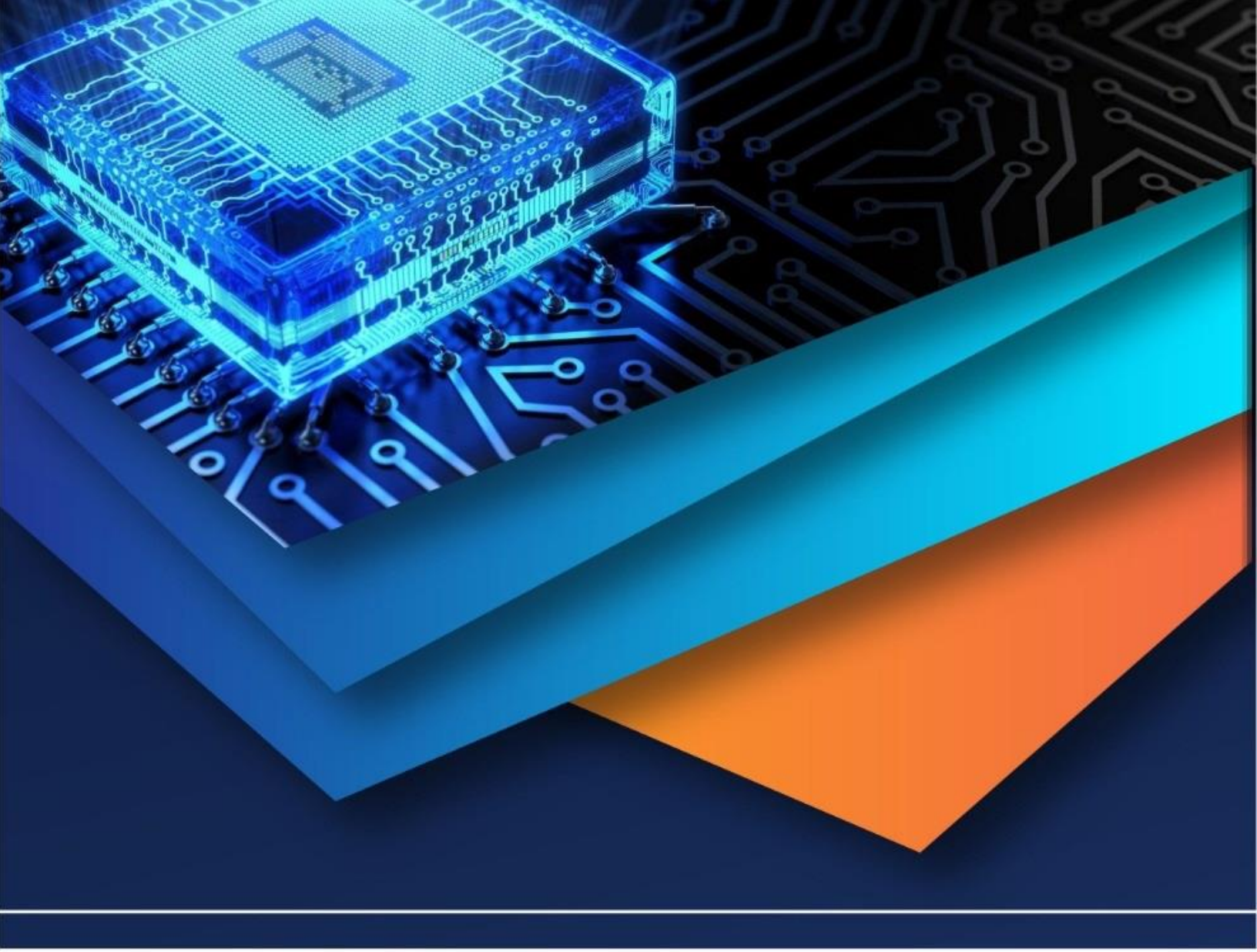
[13] Algorithms that remember: model inversion attacks and data protection law - Veale M, Binns R, Edwards L. 2018 Algorithms that remember: model inversion attacks and data protection law.Phil. Trans. R. Soc. A 376: 20180083

[14] Forgeability and Membership Inference Attacks - Zhifeng Kong, Amrita Roy Chowdhury∗, and Kamalika Chaudhuri. 2022. Forgeability and Membership Inference Attacks. In Proceedings of the 15th ACM Workshop on Artificial Intelligence and Security (AISec '22), November 11, 2022, Los Angeles, CA, USA. ACM, New York, NY, USA, 7 pages.

[15] Certifiable Machine Unlearning for Linear Models - Mahadevan, A., & Mathioudakis, M. (2021). Certifiable machine unlearning for linear models. arXiv preprint arXiv:2106.15093.

[16] Certified Data Removal from Machine Learning Models - Guo, C., Goldstein, T., Hannun, A., & Van Der Maaten, L. (2019). Certified data removal from machine learning models. arXiv preprint arXiv:1911.03030.

[17] Machine unlearning: linear filtration for logit based classifiers - Baumhauer, T., Schöttle, P., & Zeppelzauer, M. (2022). Machine unlearning: Linear filtration for logit-based classifiers. Machine Learning, 111(9), 3203-3226
[18] Delta Boost: Gradient Boosting Decision Trees with Efficient Machine Unlearning - Zhaomin Wu, Junhui Zhu, Qinbin Li, and Bingsheng He. 2023. DeltaBoost: Gradient Boosting Decision Trees with Efficient Machine Unlearning. Proc. ACM Manag. Data 1, 2, Article 168 (June 2023), 26 pages.
[19] Machine Unlearning for Random Forests - Brophy, J., & Lowd, D. (2021, July). Machine unlearning for random forests. In International Conference on Machine Learning (pp. 1092-1104). PMLR.
[20] Asynchronous Federated Unlearning - Ningxin Su, Baochun Li. "Asynchronous Federated Unlearning," in the Proceedings of IEEE INFOCOM 2023, New York Area, U.S.A., May 17–20, 2023.
[21] Machine Unlearning - Bourtoule, Lucas, et al. "Machine unlearning." 2021 IEEE Symposium on Security and Privacy (SP). IEEE, 2021.
[22] Nguyen, Q. P., Low, B. K. H., & Jaillet, P. (2020). Variational bayesian unlearning. Advances in Neural Information Processing Systems, 33, 16025-16036.



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY