Autononym: Multimodal Anonymization of Health Data using Named Entity Recognition and Structured Medical Data Processing

Hamdi Yalın Yalıç*, Murat Dörterler†, Alaettin Uçan*, Ali Yaşar Yiğit*, Adem Ali Yılmaz*
*Department of Research and Development, Tiga Information Technologies, Ankara, Türkiye
{yalin.yalic, alaettin.ucan, ali, adem}@tigahealth.com
†Department of Computer Engineering, Technology Faculty, Gazi University, Ankara, Türkiye
dorterler@gazi.edu.tr

***Abstract*—This paper presents Autononym, an AI-powered software platform capable of robustly and scalably anonymizing health data across several formats, including unstructured free-text documents, tabular datasets, and medical images in both DICOM and standard RGB formats. Autononym ensures compliance with privacy requirements such as KVKK and GDPR by using a few-shot Named Entity Recognition model for textual data. Tabular data anonymization applies formal privacy models such as k-anonymity, l-diversity, and t-closeness to generalize or suppress quasi-identifiers while preserving analytical utility. For medical imaging, DICOM files are processed through an in-house metadata scrubbing module, whereas RGB images go through optical character recognition-based detection and removal of burned-in text. The platform features a user-friendly web-based interface that facilitates both real-time and batch anonymization, designed for healthcare researchers and institutions. Technical evaluations confirm that the software is compatible with multiple input formats (csv, txt, DICOM) and has high accuracy in data anonymization. Integrated reporting modules enable the validation and auditing of data. Autononym empowers privacy-conscious health data sharing by integrating multimodal anonymization into a singular AI-driven framework. This ensures that medical research is secure and compliant.**

***Keywords*—health data anonymization, named entity recognition, medical image metadata, privacy models**

I. Introduction

The rapid digitalization of healthcare has resulted in the unparalleled accessibility of patient data through electronic health records (EHRs), medical images, and various multimodal datasets. These datasets are very useful for clinical research, AI development, and efforts to improve quality, but they also contain Protected Health Information (PHI) or Personally Identifiable Information (PII) that must be kept safe according to strict privacy laws like the Health Insurance Portability and Accountability Act (HIPAA), the General Data Protection Regulation (GDPR), and the Turkish Personal Data Protection Law (KVKK). These rules say that direct and quasi-identifiers must be removed or changed before data can be shared. This makes anonymization an important step before health data can be used again.

979-8-3315-5565-8/25/$31.00 ©2025 IEEE

Traditional manual de-identification processes, such as human review of clinical notes or manual removal of metadata, are labor-intensive, prone to error, and unsuitable for large-scale datasets. As a result, research has become more focused on automated, AI-powered anonymization techniques capable of effectively detecting and eliminating identifiers while maintaining the utility of the underlying data.

For textual health data, anonymization is mostly formulated as a Named Entity Recognition (NER) task, where models detect and mask PHI entities. While early rule-based systems struggled with generalization, machine learning classifiers, deep learning approaches and transformer-based architectures like BERT [1] have achieved state-of-the-art results. Recent models and general-purpose Large Language Models (LLMs) such as GPT-4, demonstrate high precision and recall in clinical de-identification. For medical images, anonymization involves both metadata scrubbing of sensitive Digital Imaging and Communications in Medicine (DICOM) tags and pixel-level removal of burned-in text or identifiable anatomy, often via Optical Character Recognition (OCR) and computer vision techniques. Open-source frameworks, such as those presented by Rempe et al. [2], integrate OCR-based text removal, facial feature masking, and metadata cleaning into unified pipelines, delivering competitive de-identification performance at scale.

While existing anonymization tools offer valuable functionality, most are limited to a single data modality or lack adaptability between privacy and utility. Challenges such as cross-institutional generalization, transparency, and real-time performance persist. In contrast, Autononym offers a unified, multimodal solution for anonymizing structured data, free-text, and medical images. This paper presents its architecture, methodology, and evaluation, highlighting advancements over prior approaches in AI-driven health data anonymization.

II. Related Work

Recent studies in health data anonymization cover both text and image formats. They have a focus on utilizing AI to enhance accuracy and scalability. In the field of textual de-identification, Altalla et al. [3] used LLMs, including GPT-3.5

and GPT-4, for the anonymization of clinical reports, indicating that GPT-4 achieved high precision (0.99) and robust overall performance in PHI removal tasks while preserving the contextual integrity of the text. Negash et al. [4] conducted a review of free-text PHI de-identification methods, emphasizing the growing use of hybrid approaches that integrate rule-based techniques with deep learning to improve generalizability across datasets. Ahmed et al. [5] showed that RNN models with self-attention work well for EHR de-identification, achieving a high recall rate while maintaining clinical utility metrics.

Norgeot et al. [6] developed Philter, an open-source system for anonymizing clinical text that uses a combination of methods and has been tested on one of the largest manually annotated collections of clinical notes. Moore et al. [7] introduced Transformer-DeID, a suite of transformer-based models (including BERT) for removing PHI, showing state-of-the-art performance on benchmark datasets.

In the field of medical imaging, Rempe et al. [2] introduced an open-source de-identification framework that utilizes OCR for the removal of burned-in text, metadata scrubbing, and facial feature masking, illustrating competitive accuracy and notable speed enhancements compared to previous tools. In addition, Familiar et al. [8] created a deep learning method for anonymizing facial features in magnetic resonance imaging (MRI) scans. This method successfully hides identifying anatomy while keeping structures that are important for diagnosis.

These works together show the trend toward integrated, AI-driven, and multimodal anonymization systems. This is the design philosophy behind Autononym.

III. Methodology

A. System Architecture

The Autononym system architecture, illustrated in Figure 1, is designed to handle different types of input data, such as medical images, structured datasets, and free-text documents, using specialized anonymization pipelines. A few-shot NuNER model [9] processes incoming free-text data to find and mask entities. Structured tabular data is transformed using k-anonymity, l-diversity, and t-closeness to protect quasi-identifiers. A DICOM metadata anonymizer or an OCR-based model that finds and masks burned-in identifiers can make medical images anonymous. These components work in a modular way, which means that the system can use the proper anonymization method depending on the type of data it finds. This keeps privacy compliance while ensuring that data can be used in all formats.

B. Text-based Anonymization

Autononym uses the NuNER model in a few-shot learning setting, fine-tuned with a small, domain-specific dataset customized to healthcare and Turkish language data formats and structures. This model is used for text-based anonymization to detect and hide sensitive entities without needing to fine-tune for each task. NuNER is a compact, task-specific foundation

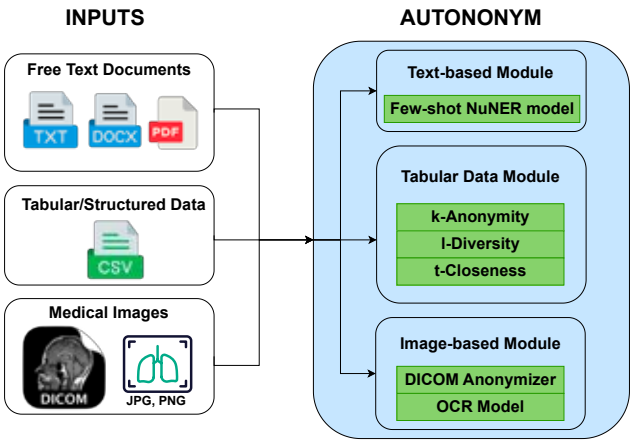

Fig. 1: Autononym system architecture

model for NER that is pre-trained on a large, diverse dataset annotated automatically by LLMs. This pre-training strategy enables NuNER to achieve strong generalization and competitive performance even with limited labeled data, allowing Autononym to handle varied healthcare datasets and formats while eliminating the dependency on manually annotated training data from the target domain.

The GLiNER framework [10] connects the underlying model and starts it up with the NuNer weights, which have been fine-tuned for strong NER across many entity categories. The system's first goal is to find a set of entity labels that represent the most common types of PII and PHI in healthcare text. These include:

$$L = \{\text{"name"},\ \text{"surname"},\ \text{"date"},\ \text{"person"},\ \text{"birthdate"},\ \text{"salary"},\ \text{"address"},\ \text{"age"},\ \text{"disease"}\}$$

Given an input text T, the model predicts a set of entities:

$$E = \{(l_i, s_i, e_i, t_i) \mid l_i \in L,\ s_i \text{ and } e_i \text{ denote start/end indices, } t_i \text{ is entity text}\}$$

where l_i is the label, s_i and e_i are character positions, and t_i is the entity string.

Adjacent entities with the same label and minimal separation ($s_{i+1} \leq e_i + 1$) are merged to form continuous entity spans. This merging step ensures that multi-word identifiers (e.g., "Dr. John Smith") are treated as a single anonymization unit rather than fragmented tokens.

Once merged, all entity spans are replaced by the masking token "$**$", producing the anonymized text T':

$$T' = \text{mask}(T, E) \quad \text{where} \quad \forall e \in E,\ T'[s_e : e_e] = \text{"}**\text{"}$$

The anonymization success rate is computed as:

$$\text{Success Rate} = \begin{cases} 100, & \text{if } |E| = 0 \\ \dfrac{|E_a|}{|E|} \times 100, & \text{otherwise} \end{cases}$$

where $|E|$ is the number of detected sensitive entities and $|E_a|$ is the number of entities successfully anonymized (in our case, all detected entities are replaced, so $|E_a| = |E|$).

This few-shot setup allows us to operate without retraining when new entity types are introduced. The model's capacity for real-time inference ensures that sensitive textual data can be anonymized on-the-fly during ingestion or before sharing, aligning with compliance requirements.

C. Image-based Anonymization

The image-based anonymization module in Autononym supports two distinct workflows depending on the format of the input file:.

1) DICOM Files: Anonymization of medical images stored in the DICOM format involves systematically removing sensitive metadata fields that could contain PHI. List of fields to remove is defined in a dictionary, such as:

$$\{\text{PatientName}, \text{PatientID}, \text{BirthDate}, \text{StudyDate}, \text{AcquisitionDate}, \text{AccessionNumber}, \text{InstitutionName}, \text{ReferringPhysicianName}, \ldots\}$$

The anonymization process starts by loading each file. This can be done with either an input file or a folder of DICOM files. Then it goes through all of the predefined sensitive fields and removes them. It also deletes private tags that might have site-specific identifiers. After that, the anonymization score is calculated using the formula:

$$\text{Score} = \frac{\#\text{ fields successfully removed}}{\#\text{ total fields to remove}} \times 100$$

Finally, the anonymized set is converted into DICOM format and encoded as a Base64 string so that it can be safely sent or stored. The output is a JSON object that has the anonymized file(s), their anonymization scores, and Base64-encoded content. This method makes sure that no metadata that could identify a patient is left behind, while keeping the original pixel data for research or diagnostic purposes.

2) RGB Image Files: If the input is an RGB image (like a .jpg or .png file), the anonymization looks for and removes burned-in text or notes that could reveal the patient's identity. To do this, OCR is used to scan the image and identify text areas.

Once these text areas are detected, the bounding boxes corresponding to these text regions are masked to prevent legibility. This method works best for modalities or export formats where identifying information is directly embedded in the visible pixels. Examples include ultrasound captures, photographic wound documentation, and exported non-DICOM radiology images.

D. Tabular Data Anonymization

For tabular datasets containing structured records, Autononym implements a privacy-preserving transformation approach that ensures the protection of sensitive and quasi-identifying attributes while keeping as much analytical utility as possible. The overall approach involves reading the input dataset, identifying quasi-identifiers, applying predefined generalization hierarchies, and enforcing selected privacy models. Suppression is used only when necessary, and its rate is monitored to maintain data usability.

Once the quasi-identifiers and their hierarchies are defined (e.g., grouping ages into ranges, generalizing salaries into classes), the system applies a set of formal privacy models to minimize the risk of re-identification:

1) k-Anonymity: Ensures that each record is indistinguishable from at least $k-1$ other records with respect to the quasi-identifiers. Formally, given a set of quasi-identifiers Q, the dataset satisfies k-anonymity if every equivalence class EC defined by Q has size $|EC| \geq k$. This prevents unique combinations of attributes from revealing individual identities.

2) l-Diversity: Addresses the weakness of k-anonymity in cases where all records in an equivalence class share the same sensitive value. The dataset satisfies l-diversity if, for each equivalence class, there are at least l "well-represented" distinct values for the sensitive attribute. One formalization is *entropy l-diversity*, where

$$-\sum_{i=1}^{m} p_i \log p_i \geq \log l$$

for each equivalence class, with p_i denoting the frequency of the i-th sensitive value.

3) t-Closeness: Ensures that the distribution of sensitive attributes in each equivalence class is close to their distribution in the overall dataset. An equivalence class satisfies t-closeness if the distance between the two distributions is no more than a threshold t. This distance is often measured using metrics such as the Earth Mover's Distance (EMD):

$$\text{EMD}(P_{EC}, P_T) \leq t$$

where P_{EC} is the distribution of the sensitive attribute within the equivalence class and P_T is the distribution of the sensitive attribute in the entire dataset.

IV. Evaluation

The figure 2 shows how to anonymize a cervical spine MRI in DICOM format. The upper-left corner of the original scan on the left has sensitive patient metadata (name, date of birth, hospital information, and acquisition details) burned into it. The medical image content remains for diagnosis, but the identifying elements have been removed from the metadata on the right. The figure 3 shows an anonymized version of a structured CSV dataset that has personal identifiers and quasi-identifiers like name, surname, national ID number, age, gender, place of birth, salary, marital status, profession, and education. The "Input Data" table shows the original values, while the "Output Data" table shows the anonymized version, where direct identifiers are replaced with masking symbols (*) and quasi-identifiers are grouped (for example, salaries are grouped into ranges and ages are grouped into ranges). The figure 4 shows an unstructured free-text document that has been made anonymous. The "Original Text" contains personal information such as names, birth dates, addresses, job details,

salaries, and information about spouses. The "Anonymized Text" keeps the same sentence structure and non-sensitive content. All sensitive fields have been replaced with masking symbols (**), so the text can still be read later.

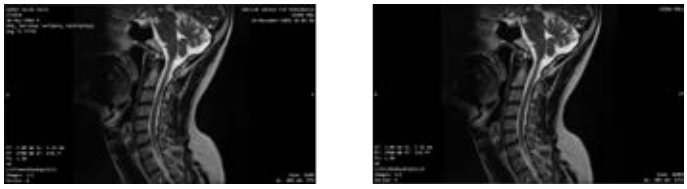

Fig. 2: Medical image (DICOM) anonymization example

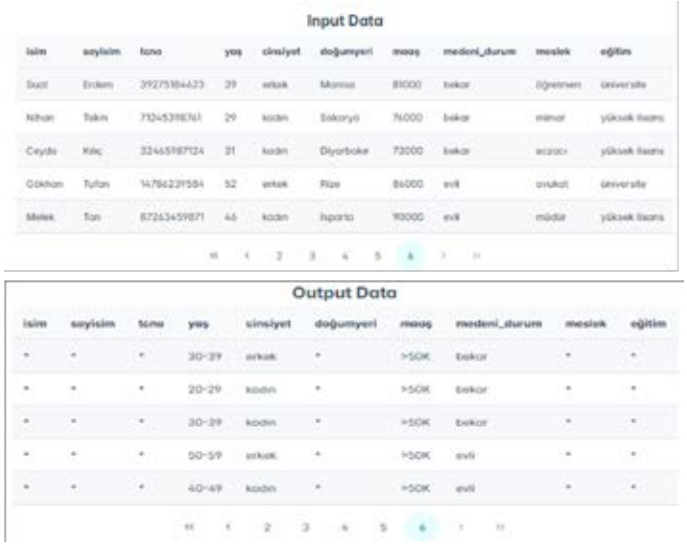

Fig. 3: Tabular data anonymization example

Fig. 4: Text data anonymization example

V. Conclusion

In this paper, we introduced Autononym, an AI-powered, multimodal health data anonymization platform designed to provide regulatory compliance while preserving data utility. The framework integrates few-shot NER for free-text entity masking, k-anonymity, l-diversity, and t-closeness models for tabular data protection, and both DICOM metadata scrubbing and OCR-based burned-in text removal for medical images in DICOM and RGB formats. It supports real-time and batch anonymization workflows by its web interface, enabling healthcare researchers and institutions to securely share data across multiple modalities. Our evaluation demonstrates high accuracy, multimodal compatibility, and flexible deployment, positioning Autononym as a comprehensive and practical solution for privacy-preserving medical research and data exchange.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, 2019, pp. 4171–4186.

[2] M. Rempe, L. Heine, C. Seibold, F. Hörst, and J. Kleesiek, "De-identification of medical imaging data: A comprehensive tool for ensuring patient privacy," *European radiology*, pp. 1–10, 2025.

[3] B. Altalla' *et al.*, "Evaluating gpt models for clinical note de-identification," *Scientific Reports*, vol. 15, no. 1, p. 3852, 2025.

[4] B. Negash *et al.*, "De-identification of free text data containing personal health information: A scoping review of reviews," *International Journal of Population Data Science*, vol. 8, no. 1, p. 2153, 2023.

[5] T. Ahmed, M. M. A. Aziz, and N. Mohammed, "De-identification of electronic health record using neural network," *Scientific reports*, vol. 10, no. 1, p. 18 600, 2020.

[6] B. Norgeot *et al.*, "Protected health information filter (philter): Accurately and securely de-identifying free-text clinical notes," *NPJ digital medicine*, vol. 3, no. 1, p. 57, 2020.

[7] C. Moore, L. Bulgarelli, T. Pollard, and A. Johnson, *Transformer-DeID: Deidentification of free-text clinical notes with transformers (version 1.0.0)*, PhysioNet. RRID: SCR_007345, https://doi.org/10.13026/7dj5-7x85, 2023.

[8] A. M. Familiar *et al.*, "Empowering data sharing in neuroscience: A deep learning deidentification method for pediatric brain mris," *American Journal of Neuroradiology*, vol. 46, no. 5, pp. 964–972, 2025.

[9] S. Bogdanov, A. Constantin, T. Bernard, B. Crabbé, and E. Bernard, "Nuner: Entity recognition encoder pre-training via llm-annotated data," *arXiv preprint arXiv:2402.15343*, 2024.

[10] U. Zaratiana, N. Tomeh, P. Holat, and T. Charnois, "Gliner: Generalist model for named entity recognition using bidirectional transformer," *arXiv preprint arXiv:2311.08526*, 2023.