JCSE Journal of Computing Science and Engineering

Regular Paper

Journal of Computing Science and Engineering
Vol.00, No.0, Jan 0000, pp. 000-000

Data Anonymization System with Integrated AI for Text Analysis and Protection of Personal Information

Eddison Steven Guillermo Benavides and Stefano Torracchi-Carrasco
Academic Unit of Informatics, Computer Science, and Technological Innovation, Research Group in Simulation, Modeling, Analysis, and Accessibility (SMA^2), Catholic University of Cuenca, 010107, Cuenca, Ecuador
eddison.guillermo.88@est.ucacue.edu.ec, jtorracchic@ucacue.edu.ec

Abstract
This study presents the development of a data anonymization system with integrated AI, which is designed to automate the anonymization of personal information in large volumes of data. In academic research, especially in areas such as medical research, privacy protection and compliance with regulations such as the General Data Protection Regulation (GDPR) are crucial. The integration of advanced Artificial Intelligence (AI) algorithms allows the system to automatically identify and remove sensitive data, such as names and personal identifiers, ensuring privacy without sacrificing information quality. The system is implemented on a robust backend using Spring Boot, which manages user authentication and security, and an intuitive frontend developed in React, which facilitates user interaction. The results obtained show excellent performance both in terms of security and usability, with fast response time even under high load conditions. This paper offers a practical and efficient solution for data anonymization in big data environments, providing an approach that can be adopted in various areas that handle sensitive information.This article is available to researchers in the GitHub repository via the following URL: [].

Category: Embedded Computing

Keywords: Data anonymization, Artificial intelligence (AI), Privacy protection, Sensitive data,Personal data protection, MD5 hashing, Spring Boot, React interface, Robust privacy.

I. INTRODUCTION

Protecting personal information is a growing challenge due to the increased use of online platforms, social networks, and large volumes of data in multiple industries, including academic research. In this context, data anonymization systems play a key role in protecting individuals' privacy by hiding or removing personally identifiable information (PII) in massive data sets.

Integrating artificial intelligence (AI) into these systems has significantly enhanced their ability to analyze and protect such information efficiently [1]. Through advanced algorithms, AI anonymization systems can process large volumes of text and automatically detect sensitive data patterns, such as names, addresses, or unique identifiers, removing them to ensure privacy [2].

For example, techniques such as k-anonymity ensure that no individual can be identified within an anonymized dataset, complying with regulations such as the General Data Protection Regulation (GDPR) in Europe[2][3].

However, the implementation of anonymizing systems has its challenges. Despite advances, AI systems may be vulnerable to re-identification attacks, which could compromise data privacy. In addition, the quality and usefulness of data for specific analyses may be affected by anonymization processes [2].

Despite technological advances, most systems still need to reach the production stage in many organizations. This is mainly due to the complexity of implementing robust systems that ensure data privacy and usability, without sacrificing security. The difficulty of ensuring adequate anonymization, coupled with the challenge of complying with the abovementioned

Open Access yy.5626/JCSE.2011.5.2.xxx http://jcse.kiise.org

Received 00 Month 2011, Accepted 00 Month 2011, Revised 00 Month 2011
* Corresponding Author

 pISSN: 1976-4677 eISSN: 2093-8020

regulations, means that many companies prefer in-house or underdeveloped solutions rather than risk re-identification and loss of privacy issues[4]. However, the system developed in this study is designed to overcome these obstacles through a secure backend in Spring Boot, allowing it to operate in productive environments with defined authentication and roles and guaranteeing secure access to anonymization functions. This solution not only facilitates regulatory compliance but also responds to the growing demands for transparency and accountability in information protection [5].

Moreover, developing an intuitive interface in React facilitates user interaction, satisfying the need for efficient and secure data management. This interface allows loading, viewing, and selecting Excel files, simplifying and anonymizing confidential information using advanced artificial intelligence techniques. Designed for medical and scientific research at a university institution in Cuenca, this tool guarantees an adaptable and effective workflow in big data environments, complementing a secure backend that reinforces regulatory compliance and protects personal information robustly and reliably.

This paper presents the development of a Data Anonymization System with Integrated AI that effectively protects personal information without compromising data quality. Its benefits, limitations, and potential impact in academic research and other areas requiring sensitive data are discussed.

II. RELATED WORK

Data anonymization systems have been widely studied due to their relevance for privacy protection in a growing digital environment. These systems seek to transform personal data so that they cannot be directly associated with individuals without additional information. According to Narayanan and Shmatikov (2008), anonymization aims to achieve data irreversibility through techniques such as pseudonymization and disassociation. These techniques make it possible to reduce the risks of identifying individuals, with the additional advantage that anonymized data are excluded from strict regulations such as the General Data Protection Regulation (GDPR) in Europe. However, it is noted that pseudonymization, unlike anonymization, can be reversible if the appropriate keys are available [6].

Several authors point out that although anonymization protects privacy, it has challenges. A recurring problem is the risk of re-identification, especially when anonymized data is combined with other sources of information. The European Data Protection Board report emphasizes that the risks of uniqueness, linkability, and inference are key in this context, highlighting the need for robust methods to mitigate these threats [7].

The system proposed in this paper differs from traditional approaches by integrating artificial intelligence techniques to strengthen anonymization. Unlike existing methods, which often focus on static data transformations, this system employs advanced algorithms to identify patterns that could lead to re-identification. This predictive capability offers a significant advantage over purely rule-based systems by addressing complex risks that emerge in large and diverse data sets. For example, while traditional techniques eliminate direct identifiers, the proposed system analyzes correlations between seemingly innocuous attributes, such as locations or habits, that could be used to infer sensitive information.

Despite these advantages, it is important to recognize the proposed system's limitations. Its reliance on machine learning models introduces potential biases arising from unrepresentative training data[8]. Furthermore, advanced computational infrastructure requirements may hinder its implementation in specific contexts. However, its ability to adapt to international regulations and its focus on robust privacy position it as a significant improvement in data anonymization.

III. METHODOLOGY

The main objective of the developed system is to guarantee the privacy and protection of personal data through advanced anonymization techniques. In a significant data context, the responsible management of sensitive information is essential for user privacy, allowing data to be used in research and development ethically and securely. This approach is essential, as information has nowadays become a valuable asset, and the ability to properly anonymize data ensures ethical use, avoiding identification risks and protecting individual confidentiality [4].

A. System Architecture

This anonymization system is designed to process confidential information from medical and scientific studies conducted at a university institution in Cuenca. The data, which includes personal and clinical details of the participants, is uploaded to the system in Excel format and must comply with strict privacy regulations, such as the General Data Protection Regulation (GDPR).

The system has a dual purpose: to protect individuals' identities while allowing researchers to perform analyses without compromising participants' privacy. The processed data is valuable for the development of deep learning models, treatment optimization, and other studies in health and applied sciences, driving safe and ethical research.

Using advanced techniques and artificial intelligence, the system automatically identifies sensitive information and applies the necessary transforma-

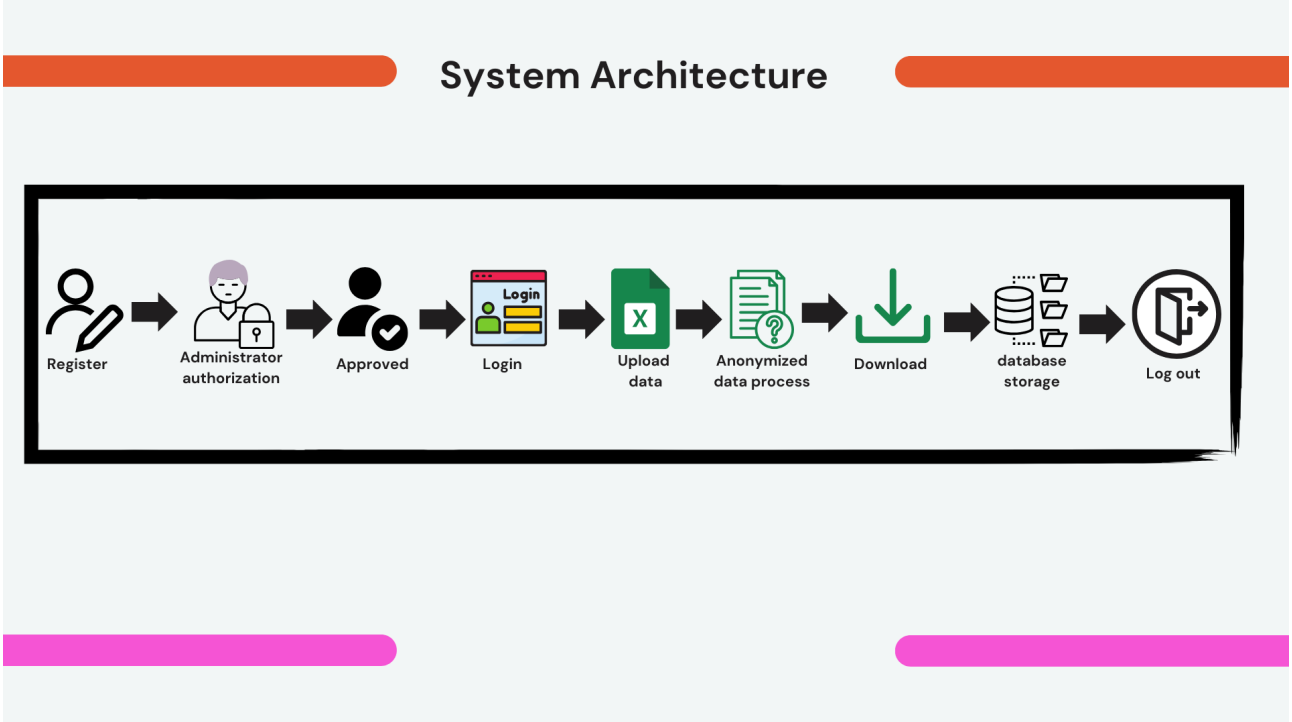

Fig. 1. System Architecture

tions to preserve the data's privacy and usefulness for future analysis.

- **Frontend (React)**

B. User authentication and security

The user interface, implemented in React, facilitates the agile and interactive capture and visualization of information. The system uses React with hooks for user authentication, which allows managing the application's state dynamically and efficiently. This methodology improves the responsiveness of the interface and allows for real-time credential capture, providing a more fluid user experience. Additionally, the system implements visual notifications through pop-up alerts (with libraries such as SweetAlert2), which provide immediate feedback to the user on their authentication status, whether success or error. Authentication is managed by a centralized service, which verifies credentials and stores an authentication token (JWT), allowing secure access to different application areas. This approach ensures that only authorized users access the corresponding information and functionalities, thus improving security and the user experience in an interactive environment.

For route protection, the ProtectedRoute and RoleGuard components are used. ProtectedRoute ensures that only authenticated users can access specific routes, redirecting unauthenticated users to the login page. In turn, RoleGuard applies role-based access control; in App.js, this component wraps specific paths, such as /users and /excelfile, allowing access only to users with the ROLEADMIN role.

Within RoleGuard, the useEffect hook verifies that the user has the JWT token and the required role via AuthService.hasRole(). If the user meets these requirements, RoleGuard allows access to the child components; otherwise, it displays an alert with Swal, redirects the user to the login page, and denies access by setting hasAccess to false. Moreover, the AuthGuard component in AuthGuard.js complements this structure by checking general authentication, redirecting unauthenticated users to the login page, and limiting access to protected paths.

C. MD5 Anonymization and AI Consumption in ExcelUpload Component

The frontend logic in the MD5 anonymization and AI consumption section for text anonymization is integrated into the ExcelUpload component. This component allows uploading and viewing Excel files, selecting specific columns to anonymize using MD5, and consuming AI services for text processing.

– **Anonymization with MD5**: The anonymize function uses CryptoJS.MD5 anonymizes specific data by transforming sensitive values into MD5 hashes. When

the user selects columns to anonymize, the handleColumnSelection function updates the data and applies anonymization to the selected cells.
- **File Upload and AI Consumption**: A file is uploaded to the backend through the handleUpload method, which returns the processed data. For AI processing, the handleIAUploadIA method sends anonymized data to the backend, where an AI applies anonymization techniques to texts and returns the updated result.
- **File Download**: Thanks to the handleDownload method, the Download Anonymized Excel button allows the user to download the anonymized file in .xlsx format..

This component combines a flexible interface with robust anonymization logic and advanced options for interacting with AI on the backend, providing a complete solution for anonymizing data in the interface, ensuring robust security management, protecting sensitive paths, and ensuring that only users with specific roles can access certain areas of the application.

- **Backend (Spring Boot)**
In the backend, developed with Spring Boot, security and authorization are configured using OAuth2 and JWT, allowing robust access control to sensitive application resources. The implementation has several key classes that structure the authentication, authorization, path protection flow, Excel file processing, and data anonymization through AI consumption.

D. Classes for Excel File Processing

The backend handles the manipulation and anonymization of Excel files through three main classes: ExcelService, HuggingFaceService, and ExcelController. These classes work together to provide complete functionality for handling Excel files and anonymizing them with artificial intelligence.

1) ExcelService:
The ExcelService class is the core of the Excel file manipulation logic in the application. Its primary responsibilities are to read, process, and save Excel files to the database and invoke AI methods to anonymize the data.

- **Method** `readExcelFile`: Este método es la entrada principal para leer archivos Excel y CSV. Detecta el tipo de archivo y lo procesa en consecuencia.
- **Method** `readWorkbook`: Processes Excel files using the Apache POI library. Allows passing an applyAI parameter to indicate whether to apply anonymization using AI.
- **Method** `writeExcelFile`: It converts the data into an Excel sheet and returns the file in byte format for download or storage.
- **Method** `saveExcelFileToDatabase`: Save the anonymized Excel file in the database.
- **Method** `getExcelFileFromDatabase`: Retrieves an Excel file from the database according to its ID.

In the anonymization process, ExcelService delegates to HuggingFaceService the responsibility of deidentifying sensitive data within the file's cells, invoking its deidentify method when necessary.

2) HuggingFaceService:
The HuggingFaceService class is responsible for integrating the AI model hosted by Hugging Face to de-identify personal data. It uses the obi/deid-roberta-i2b2 model and specializes in medical data anonymization.

- **Method** `deidentify`: This method makes a call to the Hugging Face model, sending the text to de-identify. The model returns the entities to anonymize, such as names or dates, and the method replaces that data with "xxxx" in the text.
- **Annotation** `@Retryable`: The class implements a retry strategy in case the model is temporarily inaccessible, with up to five attempts. The function stops and returns the text unchanged if it cannot access the model after several attempts.

3) ExcelController:
The ExcelController class exposes several REST endpoints to interact with file processing services:

- `/upload`: Endpoint para cargar un archivo Excel sin aplicar IA.
- `/upload-with-ia`: Endpoint que permite subir un archivo Excel y aplicar anonimización con IA al momento de la carga.
- `/download`: Endpoint to download an anonymized Excel file..
- `/list`: Endpoint protected by the role ROLE_ADMIN, which returns a list of Excel files stored in the database.

These classes work together to enable the loading, reading, anonymization, and storage of Excel files in the database, providing a complete data anonymization workflow using AI.

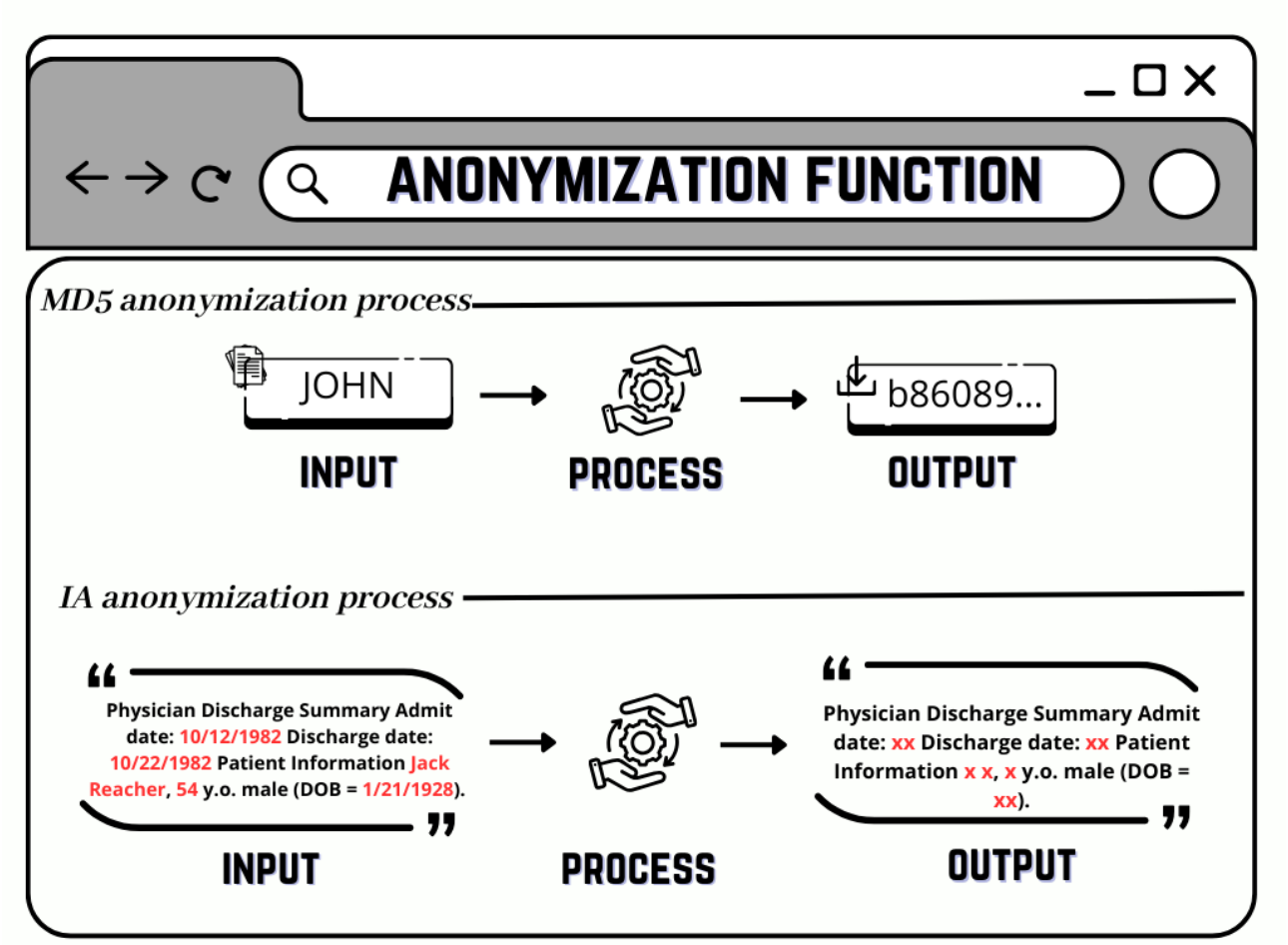

Fig. 2. Anonimyzation Function

E. Authorization Server Configuration

The AuthorizationServerConfig class is the core of the OAuth2 authorization server configuration and handles authentication and authorization. This server enables the issuance of JWT tokens for authenticated users and configures the access permissions of each registered client.

- **Customer Authentication and Token Storage**:In configure(ClientDetailsServiceConfigurer clients), the server defines the allowed clients, specifying their ID, secret, and permissions (read, write). An access token and a refresh token are assigned for each authenticated client, both stored in JWT format for reuse in future requests.
- **Authorization Endpoints**: The configure(AuthorizationServerEndpointsConfigurer endpoints) method specifies the components responsible for token management, such as the token store (TokenStore) and the JWT token converter (JwtAccessTokenConverter). In addition, InfoAditionalToken is included to add personalized information to the JWT tokens, facilitating user data consumption on the front end.
- **Security Authorization Server**: In configure(AuthorizationServerSecurityConfigurer security), it is ensured that the token validation endpoint (checkTokenAccess) is only accessible to authenticated clients. Access to the token's public key for validation on external clients is also allowed.

F. Storing and Signing JWT Tokens with `JwtConfig` y `JwtTokenStore`

The JwtConfig class defines RSA keys for signing and verifying issued JWT tokens, ensuring that only the authorized server can create and validate these tokens.

- **RSA keys for JWT**: The constants RSA_PRIVATE and RSA_PUBLIC define the private and public keys needed to sign and verify JWT tokens. The private key (RSA_PRIVATE) ensures the token's integrity when issuing it, while the

public key (RSA_PUBLIC) allows clients to validate the token's authenticity without revealing the private key.

– **JwtTokenStore Storage Configuration**: The token store (JwtTokenStore) manages issued tokens and facilitates their retrieval for subsequent validation in access requests. This approach eliminates the need to store tokens in databases, optimizing backend performance.

G. Adding Custom Information to the Token with `InfoAdicionalToken`

The InfoAditionalToken class extends the data included in the JWT token, allowing additional user information to be added to the issued token. This is done by implementing the TokenEnhancer interface, which allows customizing the content of the token before sending it to the client.

– **Token Customization**: The enhanced method uses IUsuarioService to add user data, such as name and email, to the JWT token. This information is accessible on the frontend without needing an additional request, which improves system efficiency and provides a smooth user experience.

H. Secure Password Encryption with `EncoderConfig`

The EncoderConfig class provides the BCryptPasswordEncoder password encoder, essential for storing passwords securely.

– **Using `BCryptPasswordEncoder`**: This component securely hashes passwords with a unique salt for each hash, ensuring that even identical passwords have different representations in the database. This mitigates security risks associated with password storage, such as brute force attacks.

I. Consumption of External APIs with `RestTemplate` in `config`

The config class defines a RestTemplate bean to facilitate communication with external services. This bean allows the application to make HTTP requests to third-party APIs.

– **HTTP integration with `RestTemplate`**: This component consumes external resources, such as AI services, and facilitates sending anonymized data from the front end to the back end. RestTemplate simplifies HTTP request handling by providing methods to perform operations such as GET, POST, and PUT.

Together, these backend configurations ensure secure authentication flow, sensitive path protection, and role-based access control, complementing the frontend functionality by securing the application and managing authentication efficiently.

- **Data Base (PostgreSQL)**
  The project's database is PostgreSQL, which manages and stores all application information securely and efficiently. In this context, a schema has been designed to support user, role, and file management needs, among other elements.

J. Main Tables and Relationships

In the database, the most relevant tables are:

– **Users**: This table stores the information of the system users, such as their name, email address, encrypted password, and status (active or inactive). Users are the central entity for authentication and authorization in the application.

– **Roles**: The roles table contains the different types of access available in the application, such as 'ADMIN,' 'USER,' etc. Each role has specific permissions on the paths and functionalities users can access.

– **Users_Roles**: This table is crucial for implementing the many-to-many relationship between users and roles. Each entry in this table associates a user with one or more roles. This allows a user to have multiple roles, and a role can be assigned to multiple users.

– **Archives**: The file table stores information related to files uploaded by users. These files may include data such as the file name, the upload date, and the anonymization status of the data within the file. In addition, relevant information, such as the ID of the user who performed the upload, is recorded.

K. Relations and Consultations

The database is designed to support interactions between users, roles, and files efficiently:

– The relationship between users and roles is managed through the users_roles table. This table allows multiple roles to be assigned to a user and facilitates the verification of permissions in the backend, which is essential for protecting paths according to each user's role.

– The query to obtain a user's roles is performed through a JOIN operation between the tables users and roles, using the intermediate table users_roles to combine the information.

– To manage file uploads, the system maintains traceability of which files belong to which users, allowing precise control over access and management of uploaded and anonymized data.

L. Backend Integration

The application's backend interacts with this database through SQL queries that allow authenticating users, verifying roles, managing files, and applying the corresponding security policies. By using the Spring Data JPA library in the backend, queries to the database are performed efficiently, which improves the application's scalability and performance.
For example, to authenticate a user, the backend queries the user table to verify credentials, and once authenticated, the user roles table is queried to obtain the roles associated with the user. In this way, the system can determine which routes are available based on the user's roles.
The database plays a crucial role in the application's security and administration. It ensures that only users with the correct roles can access sensitive areas and provides an organized record of uploaded and processed files. This is vital to the system's operation and to meeting anonymization and data security requirements.

M. Benefits of Anonymization and Data Security Integration in the Production Area

The relevance of integrated anonymization and data protection systems in big data environments is fundamental, as they allow the management of large volumes of information ethically and responsibly. According to Ohm (2010), anonymization is essential not only for compliance with data protection laws but also for fostering public trust in handling personal information. In addition, using anonymization techniques with robust security measures reduces the risk of re-identification, providing a secure environment that aligns with legal and ethical requirements in data handling[6].
From an operational perspective, integrating anonymization systems significantly improves efficiency in managing large volumes of data. Legal restrictions are reduced when dealing with anonymized data, allowing companies to process and analyze information more agilely without compromising privacy[6].This flexibility is especially valuable in the era of big data, where massive information analysis drives data-driven decision-making[9].
Another key benefit is increased public trust. Users are more likely to share their personal information when they perceive that their data will be treated ethically and securely. This strengthening of the relationship between consumers and companies improves corporate perception and can translate into a competitive advantage in the global marketplace [4].
Moreover, anonymization facilitates innovation by enabling organizations to perform detailed analyses and develop new products without compromising user privacy [10]. This is especially relevant in sectors such as academic research and artificial intelligence, where the availability of anonymized data can accelerate the development of advanced technological solutions[6].
Regarding security, anonymization systems are an additional barrier against fraud and malicious activity. By making it more challenging to exploit personal data, risks such as identity theft and financial fraud are mitigated, protecting individuals and organizations from significant financial loss[11].
Finally, integrating an anonymization system should be considered a long-term strategic investment. In an environment where data protection regulations are becoming increasingly stringent, and consumer expectations around privacy are constantly rising, a robust anonymization system ensures regulatory compliance and positions organizations as leaders in responsible and sustainable practices[12], [13].

IV. RESULTS

The results of evaluating the data anonymizing system with integrated AI stand out for its positive performance in key areas such as user interaction, security, and operational efficiency. In particular, the interface was rated by participants as accessible and intuitive, allowing smooth navigation through its functions. This made it easy for even non-technical users to accomplish complex tasks easily. The integrated artificial intelligence provided tremendous value by automating key processes, simplifying data anonymization, and eliminating the need for specialized technical interventions.

To illustrate how the system works, two comparative graphs are presented. Figure 5 shows an Excel file loaded before anonymization containing patient data such as name, marital status, and diagnosis. The diagnosis field includes extensive text that, in addition to describing medical conditions, incorporates personally identifiable information, representing a privacy risk. In contrast, Figure 6 displays the same file after being processed by the anonymization system. In this, single data fields, such as the name, have been transformed using the MD5 hash algorithm, ensuring their irreversibility. In addition, the integrated

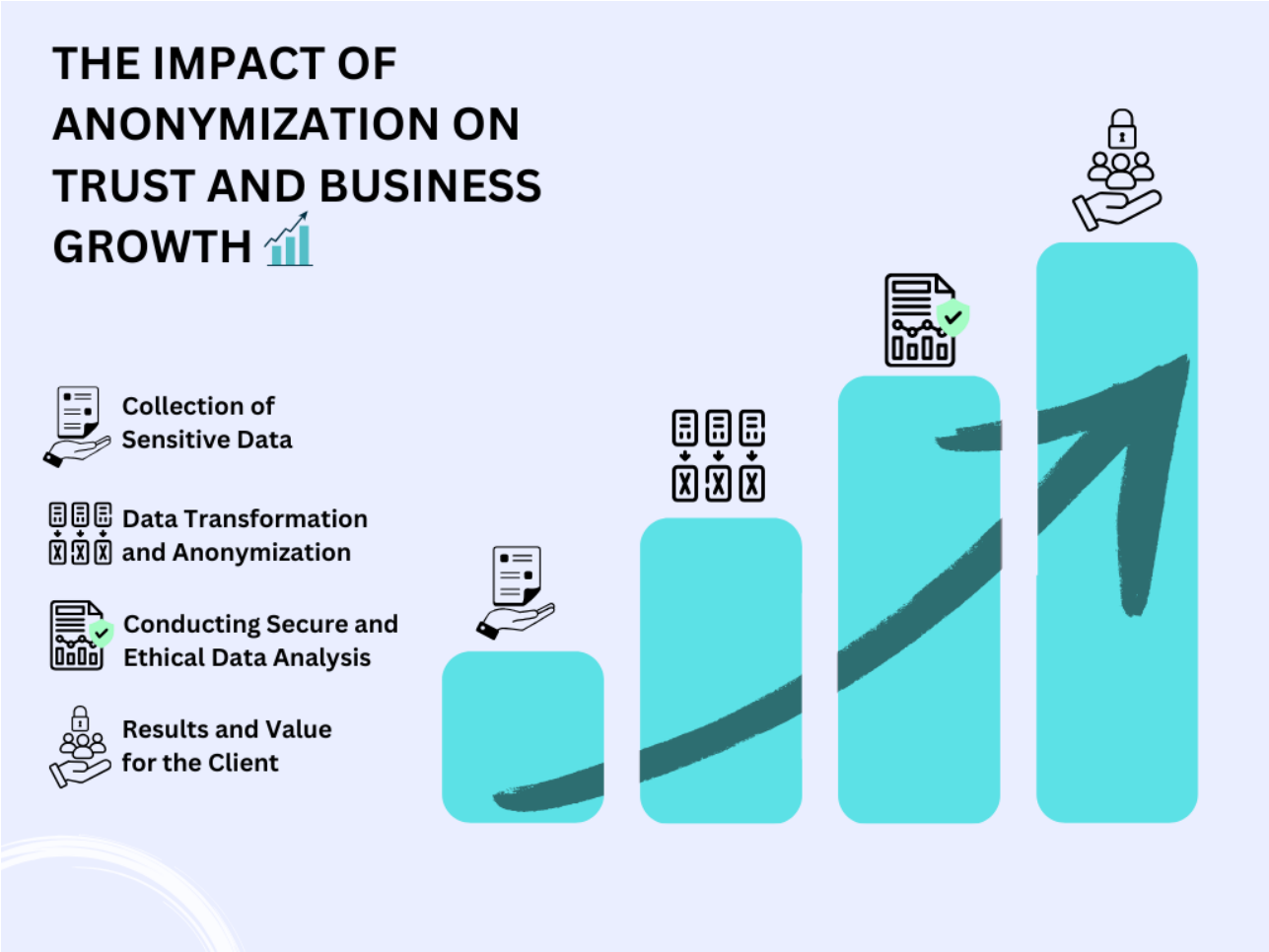

Fig. 3. The Impact of Anonymization on Trust and Business Growth

artificial intelligence automatically identified sensitive data within the diagnostic field, replacing it with the marker "xx." This hybrid approach ensures high data protection while preserving the overall file structure for further analysis.

Furthermore, the system complied with the regulatory standards established for protecting personal data, a crucial aspect in guaranteeing privacy. This regulatory compliance not only ensures the confidentiality of data but also reinforces its usefulness in the field of scientific research. Adherence to regulations such as the General Data Protection Regulation (GDPR) gives researchers the confidence to work with sensitive information ethically and legally. This creates a secure framework for analyzing large volumes of data, fostering advances in different disciplines while ensuring user privacy.

Regarding the performance of the AI model, the anonymization tests showed outstanding results. Figure 4 illustrates cases where personal entities, such as "Sarah Jessica" and "Parker," were identified with a high level of accuracy, reaching confidence values close to 1. These results reflect the system's ability to effectively protect sensitive data without affecting the integrity of the information. Additionally, stress tests confirmed the system's robustness to high volumes of requests, with consistent and fast response times. This balance between security, performance, and usability demonstrates the system's reliable and efficient solution for ensuring data privacy in dynamic environments.

V. DISCUSSION

The evaluation results highlight the advantages of the proposed AI-based data anonymization system over existing tools, such as Microsoft's Presidio [14].While Presidio offers a robust framework for anonymization of unstructured data using natural language processing (NLP) techniques, it often requires extensive configurations to accommodate highly complex datasets and specific organizational contexts. In contrast, the proposed system achieves high accuracy in identifying sensitive entities (e.g., 99.97% confidence when anonymizing personal names) and provides an intuitive user interface. This accessible design reduces technical barriers, enabling broader adoption in sectors such as research

Fig. 4. Model Test

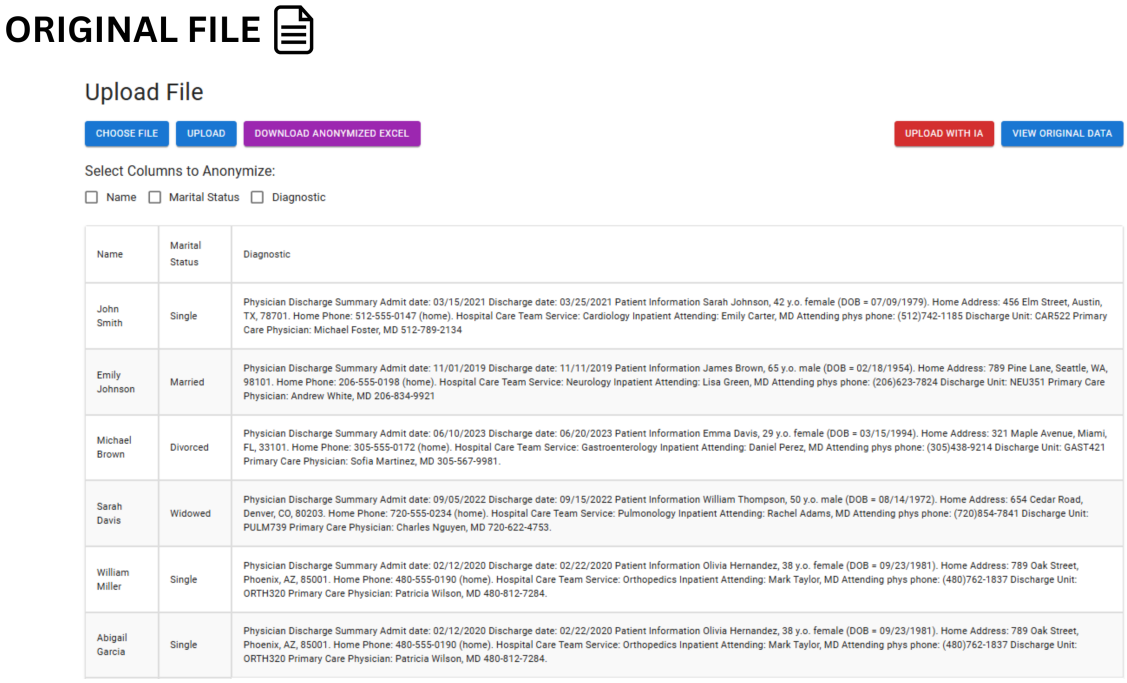

Fig. 5. Excel file before anonymization with sensitive data visible

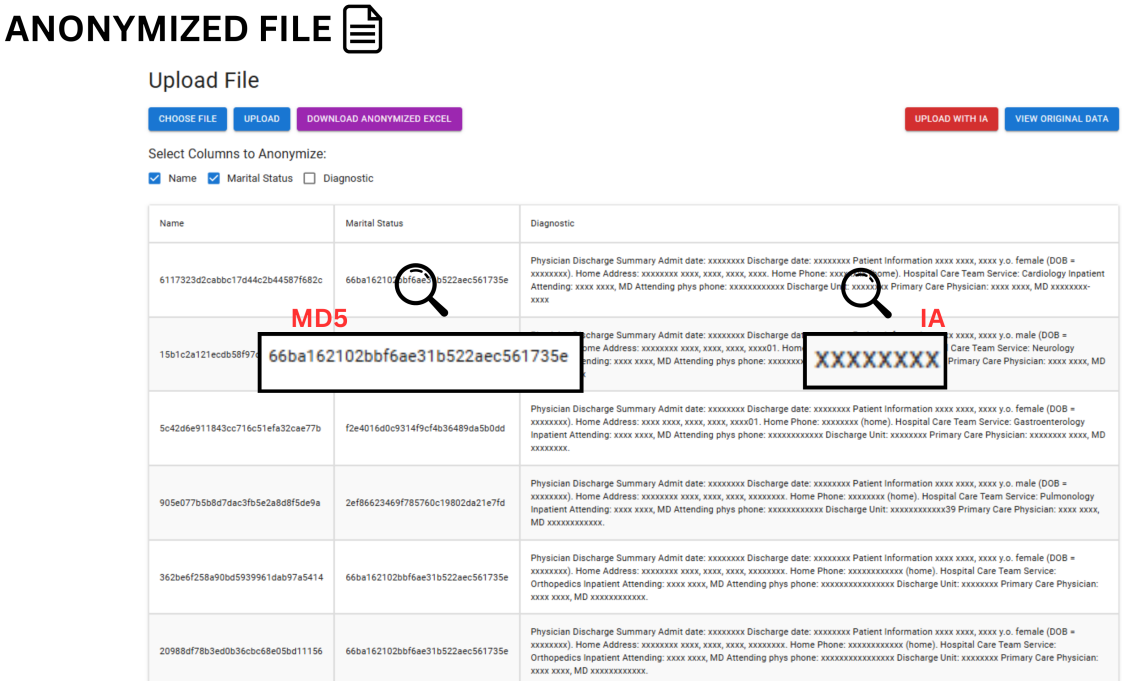

Fig. 6. Excel file after anonymization using MD5 and artificial intelligence

and healthcare while ensuring compliance with regulations such as GDPR and HIPAA. [15].

Furthermore, the developed system addresses the challenges associated with potential re-identification risks in large datasets [14]. Incorporating advanced machine learning models and multiple layers of security ensures robust privacy preservation without compromising the analytical value of the data. These features make it particularly suitable for applications in sensitive domains, such as medical and financial data management, where privacy breaches can have serious consequences.

The system also demonstrates remarkable operational efficiency, handling high volumes of data while maintaining fast response times. This performance metric is essential to ensure scalability and reliability in real-world scenarios, further cementing its suitability for diverse applications. By integrating these capabilities into a single platform, the proposed system sets a new standard for balancing usability, security, and compliance with data anonymization tools.

These advances suggest that the proposed system is well-positioned to meet the growing demand for adequate and ethical data anonymization solutions. Its focus on accessibility and robust performance ensures that organizations can confidently manage sensitive information, fostering trust among stakeholders. The development of such systems represents a critical step forward in addressing contemporary data privacy and security challenges.

VI. CONCLUSION

The data anonymization system with integrated artificial intelligence represents a significant breakthrough in sensitive data protection. A balance between ease of use and technical robustness has been achieved through a user-friendly interface and AI's ability to perform anonymization processes with high accuracy. This system ensures data privacy and promotes confidence in its implementation in complex scenarios, such as medical research or the management of large databases.

Furthermore, the evaluation shows that the system not only meets anonymization needs, but also complies with international regulatory standards, reinforcing its applicability in various legal and ethical contexts. The model's ability to identify and process sensitive data efficiently highlights its potential to revolutionize data management in key sectors, ensuring security and reliability in increasingly digitized environments.

This work lays a solid foundation for future research in AI data anonymization, highlighting the importance of integrating advanced technologies that prioritize privacy and accessibility. The possibility of expanding this technology to other languages, data domains, or high-traffic scenarios presents an exciting opportunity further to improve security and data protection in an interconnected world.

REFERENCES

[1] H.-Q. Nguyen-Son, M.-T. Tran, H. Yoshiura, N. Sonehara, and I. Echizen, "Anonymizing personal

text messages posted in online social networks and detecting disclosures of personal information," *IEICE TRANSACTIONS on Information and Systems*, vol. 98, no. 1, pp. 78–88, 2015.

[2] M. Boreale, F. Corradi, and C. Viscardi, "Relative privacy threats and learning from anonymized data," *IEEE Transactions on Information Forensics and Security*, vol. 15, pp. 1379–1393, 2019.

[3] J. S. Winter and E. Davidson, "Governance of artificial intelligence and personal health information," *Digital policy, regulation and governance*, vol. 21, no. 3, pp. 280–290, 2019.

[4] P. Ohm, "Broken promises of privacy: Responding to the surprising failure of anonymization," *UCLA law review*, vol. 57, pp. 1701–1777, 2010.

[5] D. J. Weitzner, H. Abelson, T. Berners-Lee, J. Feigenbaum, J. Hendler, and G. J. Sussman, "Information accountability," *Communications of the ACM*, vol. 51, no. 6, pp. 82–87, 2008.

[6] A. Narayanan and V. Shmatikov, "Robust de-anonymization of large sparse datasets," in *2008 IEEE Symposium on Security and Privacy (sp 2008)*. IEEE, 2008, pp. 111–125.

[7] E. D. P. Board, "Guidelines 4/2020 on the use of personal data in the context of covid-19," European Union, Tech. Rep., 2020. [Online]. Available: https://edpb.europa.eu

[8] M. Veale, R. Binns, and L. Edwards, "Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14, 2018.

[9] N. Li, T. Li, and S. Venkatasubramanian, "t-closeness: Privacy beyond k-anonymity and l-diversity," in *2007 IEEE 23rd International Conference on Data Engineering*. IEEE, 2007, pp. 106–115.

[10] J. P. Daries, J. Reich, J. Waldo, R. Diuk-Wasser, and A. D. Ho, "Privacy, anonymity, and big data in the social sciences," in *Communications of the ACM*, vol. 57, no. 9. ACM, 2014, pp. 56–63.

[11] B. C. Fung, K. Wang, A. W.-C. Fu, and P. S. Yu, "Privacy-preserving data publishing: A survey of recent developments," *ACM Computing Surveys (CSUR)*, vol. 42, no. 4, pp. 1–53, 2010.

[12] L. Taylor, L. Floridi, and B. van der Sloot, *Regulation in the Age of Anonymity*. Oxford University Press, 2017.

[13] C. Dwork, "Differential privacy," in *International Colloquium on Automata, Languages, and Programming*. Springer, 2006, pp. 1–12.

[14] I. Journals, "Data anonymization in ai and ml engineering: Balancing privacy and model performance using presidio," *IRE Journals*, vol. 6, no. 10, 2023. [Online]. Available: https://www.irejournals.com

[15] IEEE, "Data anonymization in social networks state of the art, exposure of shortcomings and discussion of new innovations," in *IEEE Conference Publication*, 2024. [Online]. Available: https://ieeexplore.ieee.org/document/9092064