Balancing Data Accessibility and Privacy: Machine Learning Approach to PII Detection in Electronic Health Records

Dissertation Manuscript

Submitted to National University

School of Technology and Engineering

In Partial Fulfillment of the

Requirements for the Degree of

DOCTOR OF PHILOSOPHY

by

ISSAH MALGOUBRI MUSAH

San Diego, California

December 2025

Abstract

This constructive research study examined the development of a scalable, context-aware machine learning (ML) framework for detecting personally identifiable information (PII) in unstructured electronic health records (EHRs). The research problem addressed the absence of reproducible, data-driven methods capable of balancing privacy preservation and data accessibility while maintaining compliance with legal frameworks such as the Health Insurance Portability and Accountability Act (HIPAA) and the General Data Protection Regulation (GDPR). The study focused on healthcare organizations and researchers who face challenges protecting sensitive health data while facilitating secure data sharing for clinical and analytical purposes.

The study's purpose was to construct, implement, and evaluate a privacy-preserving artifact guided by the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework. The research design integrated natural language processing (NLP) with unsupervised and hybrid ML algorithms, including term frequency–inverse document frequency (TF-IDF) vectorization, singular value decomposition (SVD), and density-based spatial clustering of applications with noise (DBSCAN). A transformer-based named entity recognition (NER) module utilizing Bidirectional Encoder Representations from Transformers (BERT) to validate clustering outputs. The research data were obtained from the Medical Information Mart for Intensive Care (MIMIC-III) database, a publicly available and de-identified dataset licensed through PhysioNet (Johnson et al., 2016).The experimental code and replication scripts are available at: https://github.com/NU-Academics/PII-Detection or Bert & Regular_Expression PII Detection - Colab. The model was trained and evaluated in Google Colab using BigQuery integration to ensure compliance with PhysioNet's data-use requirements.

Empirical results showed that at a sample size of 5,000 records, the model achieved a precision of 0.955 and a recall of 0.466. When scaled to 10,000 records, precision remained high at 0.854, while recall improved to 0.580. Clustering validity indices confirmed coherent separation between PII-dense and non-PII clusters (silhouette coefficient ≈ 0.38–0.45; Davies–Bouldin Index ≈ 0.95–0.99). Approximately 61 percent of the records were labeled as noise, indicating that the model effectively isolated high-risk text regions while minimizing false positives.

The study concluded that unsupervised NLP methods can reliably identify latent PII patterns within de-identified clinical narratives, achieving performance comparable to that of supervised models with lower computational costs. These findings demonstrate that scalable ML frameworks can reconcile the privacy–utility balance in EHR analytics. The research recommends incorporating hybrid explainable AI components, such as SHAP and LIME, to improve interpretability and extend future validation to institutionally governed datasets containing unredacted identifiers under Institutional Review Board (IRB) oversight.

Acknowledgments

I would not have reached this point without the grace of God, whose guidance and strength sustained me throughout this journey. I am thankful to my family for their unwavering patience, which motivated me through every challenge. I also extend my gratitude to my Subject Matter Experts (SMEs), Dr. Aeron Zentner and Dr. Amir Schur (Chair), for their constructive feedback and support throughout this research project. Finally, I would like to express my appreciation to the MIT PhysioNet Data Center team for providing the dataset used in this dissertation; their contribution was instrumental in the successful completion of this study.

Table of Contents

List of Tables

List of Figures

Chapter 1: Introduction

The Open Data initiative, established by the U.S. government, mandates that government and non-government agencies share publicly accessible datasets. Legislation such as the Open, Public, Electronic, and Necessary (OPEN) Government Data Act (Title II of the Foundations for Evidence-Based Policymaking Act of 2018) guides this initiative by outlining the standards for data collection, management, and accessibility. For example, platforms like Data.gov serve as a centralized resource for open government data, aligning with the Open Data Policy (M-13-13), which mandates transparency, accessibility, and reusability of data . However, the intersection of data accessibility and privacy protection poses significant challenges, especially when it comes to providing access to EHRs containing sensitive information. As organizations and governments increasingly adopt data transparency programs to promote data accessibility, research, innovation, and public engagement, data owners (organizations and governments) must balance these goals with the imperative to safeguard PII (Subramanian et al., 2024). The Health Insurance Portability and Accountability Act (HIPAA) define PII, referred to as protected health information (PHI), as 18 specific identifiers that can uniquely identify an individual. These include direct identifiers such as names, social security numbers, and addresses, as well as indirect identifiers like IP addresses, device identifiers, and full-face photographs. HIPAA enforces strict data privacy requirements to ensure privacy and prevent unauthorized disclosure, requiring organizations to adopt advanced privacy-preserving methods (Xu et al., 2022).

Despite regulatory frameworks such as HIPAA, traditional methods like manual review, scanning tools, and regular expressions, followed by anonymization or pseudonymization, often prove inadequate for managing large-scale and diverse data structures, particularly in EHR environments where data sharing is prevalent (Vahdati et al., 2024). According to Chang & Bergen

(2024), ML models that incorporate NLP provide scalable and practical solutions to the limitations of traditional methods for detecting PII. These models enhance data confidentiality, mitigate risks, and facilitate secure data sharing within electronic health record (EHR) systems. ML models dynamically learn and improve as new data is processed, enabling NLP to manage diverse and complex datasets effectively.

Datasets like Medical Information Mart for Intensive Care (MIMIC) are extensive, publicly available collections of de-identified health records, providing access to EHRs that include detailed patient information such as diagnoses, treatments, and free-text clinical notes. While these datasets provide valuable research opportunities, challenges are apparent in maintaining patient privacy and data security (Johnson et al., 2023): the balance between using data for research purposes and safeguarding sensitive health information strengthens the importance of developing sophisticated methods for detecting and managing PII in health data.

Models leveraging NLP can analyze unstructured data, such as free-text fields, to detect PII, especially when PII appears in non-standard formats or unexpected contexts that traditional techniques might overlook. For instance, Yu et al. (2023) proposed a clustering-based PII Model (C-PPIM) that utilizes NLP and unsupervised learning to detect PII in large unstructured text corpora. This model employs topic modeling and sequence modeling approaches to address clustering challenges in PII detection.

ML workflow for NLP can further enhance this process by labeling datasets with categories of PII, enabling more accurate and scalable detection. ML models using NLP can train on large, diverse datasets, offering suitability for extensive and heterogeneous datasets such as health records or archives. These models can adapt to new data or previously unseen patterns with minimal manual intervention or rule updates and thus improve effectiveness in identifying sensitive information. For example, a study by Chang & Bergen (2024) compared various word embeddings for biomedical NLP tasks, emphasizing the adaptability of ML models in processing complex biomedical data.

To carefully and objectively evaluate and monitor these models to ensure meaningful and unbiased clustering outcomes, assessing model performance using metrics such as the silhouette score and Davies-Bouldin index when comparing clustering results against annotated data is crucial. The selection of an appropriate clustering method depends on the specific characteristics of the dataset, including factors such as size, sparsity, and noise levels, as well as the desired clustering outcomes. For EHR datasets, clustering methods capable of handling noise, high dimensionality, and unstructured data are particularly suitable. A comparative analysis by Sarkar et al. (2024) discusses challenges and best practices for enriching digital unstructured data in health research, emphasizing the importance of selecting appropriate clustering methods for EHR data.

ML utilizing NLP identifies complex patterns and relationships within unstructured data, such as extracting insights from EHRs to predict patient outcomes, analyzing treatment effectiveness, or detecting adverse drug reactions (Johnson et al., 2023). This capability arises from ML models' ability to process large datasets and uncover hidden correlations that are challenging for traditional analytical methods. For instance, ML models using NLP can process and understand unstructured data—such as text fields in user profiles—to find and protect PII that traditional methods might overlook.

ML models leveraging NLP can continuously improve, adapt to evolving requirements, and provide scalable solutions. Continuous learning and scalability are achieved by integrating ML technologies with governance frameworks, ensuring data privacy without compromising accessibility: an approach aligns with the focus on accuracy and efficient management of large, diverse datasets. By incorporating NLP into data literacy initiatives, open EHR frameworks, and governance policies, health organizations can effectively balance the dual objectives of data accessibility and privacy.

Despite advances in ML-based privacy solutions, challenges still need to be addressed, particularly regarding the availability of high-quality training data necessary for model development. Due to the heterogeneous nature of open datasets in the context of EHRs, including variations in quality and completeness, the reliability and accuracy of ML models can be affected (Silva et al., 2022). Ensuring effective and generalizable models across different datasets is essential for broader applicability. Further research in refining and evaluating models' performance metrics and validating effectiveness is essential in research (Wu et al., 2022; Yukhno, 2024).

Statement of the Problem

The problem addressed in this study is the lack of scalable and context-aware ML models, particularly those leveraging NLP, such as transformer-based architectures, for accurately detecting PII exposure. Safeguarding PII remains a fundamental issue in data governance, especially given the increasing reliance on data-driven decision-making in healthcare systems. Although data access has expanded through digitization and open data mandates, the risk of inadvertent PII disclosure challenges ethical data reuse. Many EHR systems lack integrated mechanisms to automatically and reliably identify and classify sensitive data, particularly in unstructured formats. This absence of scalable, automated tools hinders the responsible use of EHRs in secondary research, clinical evaluation, and public health monitoring. As healthcare organizations confront evolving data privacy regulations and ethical standards, developing adaptable methods to detect PII across heterogeneous EHR data structures becomes a pressing operational and regulatory requirement.

Conventional methods, including manual reviews and essential scanning tools, are prone to errors, particularly with unstructured datasets where PII can easily escape detection. The reliance on these conventional approaches may undermine confidence in data sharing, as the dependence on human oversight to adequately identify PII creates barriers to building trust.

As data becomes increasingly diverse and proliferates across various platforms, organizations face a complex challenge in balancing the need for robust PII detection mechanisms with the demand for seamless data accessibility; this complexity comes about by the growing volume of unstructured data and the necessity to accommodate consumer expectations for data democratization (Mainetti & Elia, 2025).

Identifying and safeguarding PII is widely recognized as fundamental to maintaining privacy and enabling the extraction of valuable insights from complex datasets (Sharma et al., 2025; Negash et al., 2023). The challenges between maintaining privacy and data utility are particularly evident in sensitive domains such as mental health, where concerns about re-identification and stigma often limit the secondary use of data for research and care optimization (Im et al., 2024; Sharma et al., 2025). Addressing these challenges requires innovative approaches incorporating advanced data governance frameworks, ethical considerations, and technological solutions such as secure multi-party computation and differential privacy.

Government agencies, regulatory bodies, and cybersecurity experts, including the National Institute of Standards and Technology (NIST), the U.S. Department of Health and Human Services (HHS), and the European Union under the General Data Protection Regulation (GDPR), recommend protecting PII through anonymization and pseudonymization techniques. These methods de-identify PII by removing or replacing sensitive information with placeholders, reducing the risk of unauthorized access (NIST, 2023; Raoof, 2024). However, these techniques do not inherently detect PII categories in data or EHRs. To ensure effective de-identification, practitioners can substitute human oversight with robust ML models or complement existing techniques by using ML models to identify and classify PII before removing or masking.

A reliable ML model for detecting PII can enhance confidence in securely sharing EHRs. Accurate PII identification is critical for ensuring the proper masking of sensitive data, particularly in scenarios where human oversight is insufficient to detect PII within unstructured datasets. While data transparency and accessibility initiatives promote innovation, it also increases the risk of exposing PII.

Raoof (2024) indicated that effective detection of PII must protect sensitive data in large-scale datasets, mitigate barriers to EHR data sharing, and ensure compliance with data privacy regulations such as Health Insurance Portability and Accountability Act (HIPAA). Consequently, this approach fosters confidence in data sharing and accessibility while safeguarding personal information, particularly crucial in sectors that archive EHRs.

Conventional PII detection methods, such as manual review and rule-based scanning, lack the adaptability and speed of ML models and rely heavily on human oversight (El Azzouzi et al., 2024). Failure to address this issue may slow down the development of health research or lead to the underutilization of public health records for research purposes.

Synergizing privacy protection (through anonymization or pseudonymization) and ML within data management creates a foundation for predicting and addressing the challenges of PII detection and freedom of information requests or access (Jiang et al., 2022). Data stakeholders and researchers face increasing pressure to comply with privacy regulations while ensuring data remains accessible for analysis and innovation.

This study contributes to the field by developing a scalable ML framework for detecting PII in EHRs. The framework identifies and manages the risk of exposing personal information before data accessibility, ensuring privacy. Failure to conduct this research will perpetuate the lack of electronic health data sharing and hinder the advancement of public data transparency and accessibility initiatives, which are crucial for developing research in the health sector.

Purpose of the Study

The purpose of this constructive research study is to design, implement, and evaluate an ML-based artifact for detecting personally identifiable PII in heterogeneous EHR datasets. The study addresses the challenge of balancing data accessibility with privacy protection by creating a practical, context-aware solution that directly supports responsible data governance and enhances compliance with legal frameworks such as HIPAA and the EU AI Act.

As open data initiatives expand, efforts to promote transparency and public engagement through data sharing have concurrently increased the risk of privacy violations, particularly within regulatory frameworks such as the HIPAA. In 2023, the U.S. Department of Health and Human Services (HHS) Office for Civil Rights (OCR) reported 553 large-scale healthcare data breaches, affecting over 109 million patients. Similarly, in 2022, HHS reported that hacking and IT incidents accounted for 74% of breaches involving 500 or more individuals, impacting over 32 million people. These incidents show the growing challenges in balancing open data practices with stringent privacy protections (Shojaei et al., 2024). Traditional PII detection methods, such as manual review, rule-based scanning, and Regular Expressions (RegEx), often used alongside anonymization and pseudonymization, need to be enhanced or replaced with ML-based approaches for more effective PII detection.

By employing a constructive research methodology grounded in ML, this study aims to provide an evidence-based assessment of the efficacy of ML techniques in detecting PII and ensuring compliance with HIPAA while supporting data accessibility for analysis, innovation, and governance. Previous research has shown that traditional rule-based and manual approaches lack scalability and adaptability, reinforcing the need for advanced ML models to enable secure and accessible data sharing (Jiang et al., 2022) The outcomes of this study inform the development of scalable, adaptable, and privacy-preserving models essential for the sustainability of open data initiatives.

Introduction to Theoretical or Conceptual Framework

Ensuring data privacy while maintaining accessibility requires balancing data sharing with compliance with PII security regulations, such as HIPAA. Integrating NLP-based ML models into data pipelines enhances the detection of sensitive data categories. This approach can prevent inadvertently sharing PII by ensuring appropriate protective measures aligning with data privacy and security standards. Clear privacy policies and adaptive frameworks help mitigate costly breaches and comply with evolving data and regulatory requirements (Schäfer et al., 2023).

Regulations influence data collection, storage, and use, shaping access while preserving privacy. While necessary, stringent privacy controls may limit strategic data use. Managing tensions between data democratization and privacy protection is vital for fostering transparency and innovation (Subramanian et al., 2024). Reliable anonymization and pseudonymization reduce the risk of exposing sensitive records while balancing data access and privacy (Vahdati et al., 2024).

This research aims to build and evaluate an ML model leveraging NLP to identify PII and apply unsupervised learning techniques for clustering and dimensionality reduction. Ineffective privacy strategies can result in serious economic consequences. Innovative governance models that combine ML-based PII detection directly relate to organizations' efforts in managing sensitive data within open data ecosystems. These models enable institutions to maintain compliance with legal frameworks such as HIPAA while harnessing data for research, policy development, and service improvement. By incorporating privacy-preserving technologies into governance structures, organizations can mitigate the risk of data breaches by making data widely accessible for economic and social value. One practical example is Microsoft 365 E5; it integrates ML-driven compliance solutions—including data loss prevention, information protection, and risk-based access controls—into its cloud governance infrastructure to support secure and compliant data use, an approach that promotes responsible data sharing, fosters public trust, and ensures that data-driven initiatives adhere to ethical and regulatory standards (Appiah et al., 2021; Soini et al., 2022).

ML approaches improve the detection of PII by identifying context-sensitive attributes that traditional methods, such as manual review and Regular Expressions (RegEx), frequently miss. RegEx relies on deterministic pattern matching, which makes it inadequate for detecting obfuscated or irregular identifiers that do not follow fixed syntactic structures (Lin et al., 2025). Balancing accessibility and privacy is especially critical for large-scale public datasets. Current models, such as those using differential privacy, often struggle, underscoring the need for governance structures integrating ML to manage these risks (Yukhno, 2024). Differential privacy adds controlled noise to sensitive data, offering a structured method to balance utility and privacy (Xu et al., 2022).

Information Theory provides the core theoretical foundation for this study by framing PII detection as a problem of reducing uncertainty in textual data. Entropy, defined as the expected level of randomness or disorder in a dataset, quantifies the unpredictability or ambiguity of tokens within clinical narratives. Higher entropy typically signals or reflects irregular, context-dependent, or weakly structured identifiers, such as nonstandard names, contact references, or implicit occupational clues (Zhang et al., 2025). Clustering methods, such as DBSCAN and K-Means, inherently minimize entropy by identifying dense regions of similar representations and isolating high-risk noise regions, which often correlate with hidden or unstructured PII (Wang et al., 2020). Likewise, transformer-based models, such as BERT, reduce entropy through contextual embedding, assigning probabilistic meaning to entities based on surrounding text (Chung, 2023). This theoretical framing directly aligns with RQ2 and RQ5, which examine latent pattern discovery and risk quantification in PII detection.

Although VC Dimension is traditionally used to evaluate model capacity and generalization bounds, this dissertation does not calculate VC Dimension empirically. Instead, it conceptually acknowledges that model selection must align with the structure of heterogeneous healthcare data, where irregular PII formats vary in complexity. The theoretical implication is that high-capacity models, such as transformer-based architectures, are better suited to capture contextual dependencies for identifying irregular PII, without mathematically deriving VC bounds (Zhu et al., 2023).

This study applies foundational concepts from ML theory to develop a scalable and privacy-conscious framework for detecting PII in EHRs. Recent advancements reinterpret the NFLT, suggesting that real-world data often exhibit low Kolmogorov complexity, enabling neural networks to generalize effectively across diverse tasks without necessitating task-specific architectures (Liu et al., 2024). This perspective justifies the exploration of multiple unsupervised learning algorithms—such as k-means, DBSCAN, and hierarchical clustering—to identify the most effective model for PII detection in varied EHR datasets.

Information Theory provides the analytical foundation for this study, where entropy quantifies uncertainty in identifying PII within heterogeneous clinical narratives. In text-based PII detection, higher entropy often reflects ambiguous or context-dependent identifiers, which clustering and transformer-based NLP models help reduce through probabilistic and contextual inference (Alvarez-Melis & Jaakkola, 2023; Li & Sun, 2024).

Iterative refinement, a common practice in unsupervised learning, involves tuning parameters (e.g., number of clusters, distance thresholds) and assessing performance using validity indices such as the silhouette score and Davies–Bouldin index, facilitating model convergence and performance optimization.

Integrating ML models with privacy regulations, including the GDPR and the HIPAA, necessitates understanding the trade-offs between model accuracy and data minimization. The generalization's performance directly informs these trade-offs.

Introduction to Research Methodology and Design (Nature of the Study)

The study employed a constructive research design, which focuses on creating artifacts, tools, or models as central elements of the research process while systematically documenting and validating the associated methods and knowledge. Although a historical challenge in constructive design research has been the struggle to bridge the gap between practical applications and academic literature or discussions, a new phase of the constructive design framework leverages domain-specific skills and expertise (Delgado Ramos, 2023). Constructive design emphasizes creating and evaluating tangible artifacts—such as computational models—as a means of generating knowledge. In this context, this research develops and assesses an ML framework for detecting PII in EHRs, integrating technical implementation with theoretical and empirical justification to address privacy concerns while supporting secure data access.

A process model is created alongside a constructive methodology to explain the steps involved in constructing the ML framework for detecting PII. Traditional data detection methods often face scalability, adaptability, and accuracy challenges when processing large-scale or unstructured datasets, making efficiency more challenging to validate than ML leveraging NLP, which has demonstrated superior performance in handling such data complexities. The building of this model or tool aims to ensure the reliable identification of PII records (Kohli et al., 2025).

Data is collected from publicly available EHR repositories or archives for processing using NLP and text analytics to extract relevant information. The data NLP techniques enable large-scale processing, essential for identifying PII and training and evaluating ML models effectively. The study employs a constructive design approach to develop a tool that identifies variables such as types of PII, ensures regulatory compliance, and evaluates the accuracy of the ML framework. This non-causal approach provides insights into the factors influencing the model's effectiveness in achieving dual objectives (Delgado Ramos, 2023).

This research ensures appropriate representation within the dataset by employing a stratified sampling approach using publicly available data from PhysioNet. Although the dataset originates from a single repository, it includes multiple clinical sources, patient demographics, and types of medical records. The stratified sampling preserve variation across available dimensions, such as patient age groups, clinical conditions, record types (e.g., discharge summaries, ECG reports), and documentation practices, ensuring that the ML model is trained and evaluated on a dataset reflecting the heterogeneity within PhysioNet's corpus. By including diverse documentation contexts and potential PII exposure patterns, the approach supports a more reliable assessment of the model's generalizability and robustness in detecting PII across varied clinical narratives. For initial exploration analysis, the use single-source datasets to provide a controlled environment for focused testing and iterative refinement. The study prioritizes depth and iterative development over broad generalizability at the early stages, aligning with constructive research principles (Peffers et al., 2020).

The use of constructive research methodology emphasizes iterative refinement in knowledge development through artifact construction rather than pursuing broad generalizability in early phases. The research enhances internal validity by applying NLP techniques to reduce human error in identifying and classifying PII. In assessing external validity, the ML model must be tested on multiple subsets of the PhysioNet dataset, examining its performance across different clinical narratives and patient demographics. This study also apply abductive reasoning to iteratively refine hypotheses based on empirical findings iteratively, aligning with the principles of constructive design in which experimentation and insight develop concurrently.

Research Questions

This study's research questions aim to investigate how unsupervised ML models, particularly those using NLP and clustering techniques, can optimize the balance between data accessibility and privacy compliance in EHRs containing PII. The following outlines the key challenges that this study seeks to address.

RQ1

To what degree can ML models integrating NLP enhance the precision of PII detection in EHRs?

RQ2

To what extent can ML models employing unsupervised NLP techniques uncover specific, previously undetected PII patterns, such as irregular data formats or contextual identifiers, within unstructured EHR datasets?

RQ3

How do NLP-based methods compare to manual validation in effectiveness (precision, recall) and scalability (processing time) for PII detection in large-scale EHR datasets?

RQ4

To what extent do NLP-based PII detection techniques improve HIPAA-aligned de-identification by optimizing the trade-off between data privacy (e.g., precision and recall of PII detection) and data accessibility (e.g., preservation of clinical utility) in EHR?

RQ5

How can NLP-based PII detection be used to quantify potential data exposure risks within annotated EHRs?

Significance of the Study

This dissertation addresses a persistent and intensifying challenge in health informatics: ensuring the accessibility of EHRs for research and innovation while maintaining stringent protection of PII. In 2025, this challenge remains central to the development of ethically responsible data infrastructures, particularly as healthcare systems face increasing pressure to support real-time analytics, AI-driven clinical decision tools, and multi-institutional research networks (Han et al., 2022; Ambati, 2025) While foundational to advancements in personalized medicine and public health surveillance, the widespread digitization of health data has concurrently heightened exposure to privacy violations and re-identification risks, prompting stricter compliance expectations under regulatory regimes such as the HIPAA in the United States and the GDPR in the European Union.

This study presents an ML-based framework integrated with NLP to automate PII detection in unstructured EHRs. This approach offers a flexible and scalable alternative to conventional anonymization or de-identification techniques, which frequently rely on rule-based methods and deterministic redaction mechanisms. Researchers have widely implemented traditional techniques, but these conventional methods often compromise the analytical fidelity of data and leave it vulnerable to re-identification attacks, particularly in narrative clinical notes, discharge summaries, and other semi-structured text sources (Tertulino et al., 2024). By employing transformer-based NLP architecture, the proposed method enhances the detection of context-dependent identifiers, such as occupation, rare conditions, and institutional affiliations—without degrading downstream data usability. Researchers comply with privacy regulations and retain the analyzed datasets critical semantic and statistical properties through this approach (Kuo et al., 2025).

The practical significance of this research extends to institutions and stakeholders operating under financial or infrastructural constraints. Proprietary tools, such as Microsoft Purview and AWS Macie, offer advanced data protection features. However, the licensing costs, vendor lock-in, and black-box design limit transparency, preventing researchers from auditing why specific identifiers were flagged or redacted and inhibiting reproducibility in privacy research (Liu & Sun, 2024). This study designs the framework as a low-cost, platform-independent utility that researchers can locally deploy or integrate into secure environments, and this increases equitable access to PII detection and compliance infrastructure, aligning with ongoing global initiatives to democratize privacy-preserving technologies across healthcare systems (Weiner et al., 2024; Krishnamoorthy, 2025).

In addition to financial constraints, commercial PII detection tools such as AWS Macie and Microsoft Purview operate as proprietary "black box" systems that restrict transparency and internal auditability. These systems typically do not disclose how redactions or classifications were generated, preventing researchers from tracing model decisions, validating false positives or negatives, or examining how contextually sensitive identifiers were interpreted (Wang et al., 2023). This limitation presents practical and ethical challenges in clinical and research environments, where transparency, reproducibility, and verifiable reasoning are essential for regulatory compliance and methodological integrity (Liu & Sun, 2024). By contrast, the unsupervised and open-source framework developed in this study facilitates auditability by allowing researchers to trace cluster-level decisions, inspect contextual embeddings, and validate entity extraction steps. This transparency supports explainable AI (XAI) approaches, ensuring the model's decisions can be reviewed, interpreted, and improved based on empirical analysis rather than proprietary constraints (Liao et al., 2024).

This study engages in emerging methodological and ethical debates surrounding using artificial intelligence in clinical contexts. Despite the growing efficacy of ML systems in tasks such as image classification, disease prediction, and text analysis, concerns regarding algorithmic bias, transparency, and fairness persist (Tölle et al., 2025; Kovačević et al., 2024). ML models trained on homogeneous or poorly curated datasets can introduce inequities by underperforming across demographic subgroups or misclassifying underrepresented entities in sensitive fields such as PII detection. This research addresses these concerns by incorporating fairness-aware design principles and validating model performance across demographically diverse datasets. It contributes to developing robust, explainable, and ethically sound data processing pipelines for biomedical applications (Brodersen et al., 2025).

This dissertation responds to an important contemporary problem in digital health governance by advancing technical innovation, policy relevance, and equitable access. Integrating state-of-the-art ML and NLP for privacy-preserving data sharing strengthens the theoretical foundation of informatics. It provides practical value to institutions seeking to balance innovation with compliance and ethical responsibility.

Definitions of Key Terms

The following are key terms used throughout this dissertation. This compilation enhances conceptual clarity and ensures consistency in terminology.

Anonymization

Anonymization involves modifying or removing personal identifiers in datasets to prevent the re-identification of individuals. This method allows researchers to use health records while maintaining patient privacy (Vahdati et al., 2024).

Data Governance

Data governance is the establishment of policies and procedures to manage data quality, security, and compliance with regulations. It is essential in publicly shared health records to ensure privacy and accessibility are balanced (Subramanian et al., 2024).

Data Privacy

Data privacy encompasses strategies and regulations to protect personal data from unauthorized access. In health data sharing, privacy measures help organizations comply with frameworks such as HIPAA (Silva et al., 2022).

Differential Privacy

Differential privacy is a framework that uses mathematical techniques to ensure the privacy of individual data points by making slight alterations. This approach supports large-scale data analysis without risking individual identification (Xu et al., 2022).

EHRs

EHRs are digital versions of patient records that include medical history and personal information. These records support healthcare research and patient care but pose privacy concerns, particularly regarding handling identifiable information (Yoon et al., 2023).

ML

ML algorithm learns and improves by identifying data patterns without being explicitly programmed. ML can be applied to detect patterns, such as PII, within large datasets (Xu et al., 2022).

NER

NER is an AI-based technique that identifies and categorizes entities such as names and organizations within text data. It is valuable for detecting PII in clinical notes and other health records (Delgado Ramos, 2023).

NLP

NLP enables computers to analyze human language. Within EHR, NLP helps identify and extract sensitive data, such as patient information in unstructured text, by analyzing textual content (Delgado Ramos, 2023).

PII

PII includes any data that can identify an individual, such as names or social security numbers. Protecting PII in EHRs is necessary for compliance with privacy laws while supporting research initiatives (Wu et al., 2022).

Privacy-preserving Data Publishing

Privacy-preserving data publishing includes techniques that allow data sharing for research while safeguarding sensitive information, aligning with responsible EHR data usage (Silva et al., 2022).

Pseudonymization

Pseudonymization replaces private data with pseudonyms, allowing data to be re-identified under certain conditions when needed, thus maintaining data linkage in secure settings (Ahmed et al., 2020).

Summary

This dissertation investigates the detection of PII in EHRs, addressing the evolving challenge of safeguarding patient privacy while maintaining data accessibility for research, policy evaluation, and clinical innovation. With the expansion of artificial intelligence (AI) in biomedical informatics, healthcare systems in 2025 face increasing pressure to reconcile data utility with stringent privacy regulations. As recent reports show, many healthcare organizations continue to experience privacy breaches—many stemming from insufficient PII detection protocols in unstructured clinical data (Conduah et al., 2025; U.S. Department of Health and Human Services, 2024). A constructive methodology aligned with the CRISP-DM framework (Ma et al., 2025) facilitates the iterative development, evaluation, and refinement of ML-enabled automated solutions that classify PII within diverse, large-scale EHR datasets. This methodology circumvents the resource demands of traditional NER models, which depend on manually annotated datasets. Instead, unsupervised learning algorithms—such as k-means and DBSCAN (Density-Based Spatial Clustering of Applications with Noise)—can identify clusters of similar data points based on density and isolate rare or anomalous entries as outliers. DBSCAN is particularly suited for detecting uncommon and potentially sensitive information embedded in complex clinical text.

DBSCAN When combined with transformer-based NLP models, such as BERT (Bidirectional Encoder Representations from Transformers), which captures the bidirectional contextual meaning, and Roberta (Robustly Optimized BERT Pretraining Approach), which enhances BERT's architecture for improved performance, the approach enables context-sensitive detection of both direct identifiers (e.g., names, Social Security numbers) and indirect identifiers (e.g., rare diseases, job titles). These transformer models generate deep semantic embeddings of clinical text, capturing nuanced term relationships. Applying DBSCAN to these embeddings allows for identifying common and outlier patterns that may signify the presence of PII. This integration enhances detection precision while maintaining the analytical utility of the underlying data (Ahmed et al., 2020; Sarkar et al., 2024).

In addition to its technical contributions, the study foregrounds regulatory compliance by aligning its model architecture with the HIPAA 18-identifier standard and GDPR's data minimization and purpose limitation principles. The framework supports institutional efforts to audit, classify, and mitigate potential privacy risks before data sharing, particularly in multi-institutional research collaborations. By integrating stratified data sampling, bias detection routines, and fairness-aware classification techniques, the research also advances ethical and equitable data science practices in line with current global guidelines for responsible AI (Kovačević et al., 2024; Weiner et al., 2024).

Moreover, this research contributes to closing the accessibility gap in privacy-preserving technology. While technically robust, commercial tools such as Microsoft Purview and AWS Macie remain financially inaccessible for smaller healthcare institutions, community clinics, and non-profit research centers. The open-source nature of the proposed framework supports its deployment as a cost-effective, platform-independent utility, enabling broader adoption across resource-constrained settings (Tölle et al., 2025); this is especially pertinent in low- and middle-income countries, where scalable privacy infrastructure remains limited despite growing data collection and analysis demands (Krishnamoorthy, 2025).

In conclusion, this dissertation contributes to both theoretical and applied domains of health informatics by offering a transparent, efficient, and ethically aligned framework for automated PII detection in EHRs. It strengthens the foundation for future privacy-preserving infrastructures, empowering data stewards to meet the dual objectives of patient privacy and data-driven discovery. Subsequent chapters elaborate on the research questions, literature foundations, methodology, and empirical findings necessary to validate this framework's utility in real-world health data environments. This study does not propose full automation of PII detection but instead introduces a triage-based ML architecture designed to reduce processing burden through prioritization and noise isolation.

Chapter 2: Literature Review

The growing reliance on EHRs for medical research, patient care, and policy development has heightened concerns about data privacy and accessibility, as healthcare datasets increasingly expose sensitive personal information despite de-identification efforts (Shojaei et al., 2024). The data domain for this study is healthcare, specifically EHRs, with heterogeneity in characteristics, volume, and sensitivity. EHRs often include free-text clinical notes, structured data, and imaging, making PII detection a multifaceted problem. The integration of ML into this domain has shown promise in improving the accuracy and efficiency of PII detection, thereby enhancing data privacy without compromising accessibility (Altalla' et al., 2025). Reviews of relevant literature on data science methodologies, ethical considerations, and legal frameworks governing PII detection support this chapter.

Limitations of traditional PII detection methods, which rely on manual scanning tools, records, and human oversight, are often overwhelmed by the scale and complexity of modern datasets. These methods are prone to errors when handling large datasets due to limitations in human capacity and the increasing volume of data, and often cannot adapt to heterogeneous EHR formats, raising concerns about complete de-identification (Chamikara et al., 2020; Hong et al., 2024). In today's data-driven world, how data is collected, processed, and managed has become increasingly sophisticated and diverse, necessitating more advanced solutions (Chen et al., 2020).

The limitations of conventional PII detection methods highlight the need for ML tools that can process large datasets with incredible speed and comparable reliability (Kocaman et al., 2023). Data science methodologies, particularly ML, have emerged as powerful tools for addressing this challenge. ML models can be adapted to address the dynamic and diverse challenges of the modern data landscape, ensuring compliance with evolving data privacy regulations such as the GDPR and the California Consumer Privacy Act (CCPA) (Rieke et al., 2020). Organizations can leverage ML capabilities to enhance PII detection processes, reduce errors, and maintain robust data privacy standards.

Ensuring data privacy while maintaining accessibility in a data management framework requires carefully protecting sensitive information and enabling appropriate data access or facilitating data usability (e.g., balancing patient confidentiality with open data access for research and policy development). Although challenges exist when managing competing interests, advanced privacy models and techniques offer promising solutions. Although challenges exist when managing competing interests, advanced privacy models and techniques offer promising solutions. Ensuring PII data privacy in data pipeline processes is vital as organizations seek to leverage data strategically while ensuring robust protections. Protecting sensitive data can help organizations avoid or mitigate costly breaches of privacy laws, achievable through clear policies that respect privacy and adaptive frameworks that respond to changes in data volume and regulatory requirements (Schäfer et al., 2023).

Regulations influence how data is collected, stored, and used, impacting access while preserving privacy. Schäfer et al. (2023) emphasize that stringent privacy controls may limit strategic data use; however, careful management of data privacy, including the protection of PII, enables transparency and innovation while making data accessible to qualified users, also referred to as data democratization (Subramanian et al., 2024). Vahdati et al. (2024) review techniques like anonymization to prevent public exposure to sensitive data, addressing the balance between data access and privacy. Additionally, pseudonymization can further reduce the risk of exposing sensitive records.

This literature review sources peer-reviewed articles and conference proceedings from multiple academic databases and search engines, including PubMed, IEEE Xplore, the ACM Digital Library, and Google Scholar. Keywords include: "ML," "PII detection," "EHRs," "data privacy," "data accessibility," "federated learning," "BERT," and "CRISP-DM." The review considers articles published between 2019 and 2025. After applying relevance and quality screening criteria, the search yielded approximately 29 articles.

Theoretical Framework

This section presents the theoretical frameworks that underpin this research, grounded in ML methodologies and privacy-preserving approaches to address the challenges of PII detection in EHRs systematically.ML frameworks, particularly transformer-based models (e.g., BERT) and Federated Learning (FL), offer a foundational approach to addressing regulatory requirements while maintaining data utility in EHRs (Tölle et al., 2025; Liu et al., 2023). Despite protecting raw data, FL models encounter challenges such as model drift, communication overhead, and uneven data distribution, requiring continuous optimization (Kuo et al., 2025). Transformer models, which utilize attention mechanisms, have significantly enhanced NLP tasks by enabling context-aware text analysis, particularly relevant to PII detection in EHRs, where distinguishing between sensitive and non-sensitive information requires contextual understanding. Despite protecting raw data, FL models encounter challenges such as model drift, communication overhead, and uneven data distribution, requiring continuous optimization (Kuo et al., 2025).

While classical ML theory often employs VC Dimension to quantify generalization performance, the present study does not compute VC Dimension values or evaluate classifiers using VC theoretical bounds. Instead, this research focuses on entropy-driven uncertainty reduction and contextual inference, which align more closely with unsupervised cluster discovery and transformer-driven contextual NER. This theoretical positioning aligns with recent perspectives that argue capacity measures, such as VC Dimension, are less informative for modern deep learning applications in unstructured text analysis, particularly in PII detection (Sokolic et al., 2024).

Entropy measures uncertainty in feature distributions and is widely used in NLP to evaluate both uncertainty and deviation from expected language patterns. In PII detection, high-entropy regions often contain irregular identifiers such as initials, shorthand phone numbers, or context-dependent mentions of individuals, which are difficult to detect using rule-based methods (Zhang et al., 2024). Clustering algorithms such as DBSCAN operationalize entropy by identifying sparse, high-variability regions that frequently represent anomalous or sensitive information (Wang & Zhang, 2023). These noise-labeled regions align with the research findings, which indicate that approximately 61 percent of extracted notes contain potential exposure risk zones. Furthermore, transformer-based models reduce entropy through attention-driven embedding representations, mathematically optimizing for contextual certainty around entities, which supports the use for both direct and indirect PII (Li & Sun, 2024). This framing links directly to research questions RQ2 and RQ5 by explaining how entropy reduction facilitates latent pattern detection and privacy risk scoring.

Constructive research is a design-oriented methodology emphasizing the development and practical evaluation of technological artifacts. In data science, it supports the creation of novel solutions—such as algorithms, predictive models, and intelligent systems—by integrating theoretical foundations with experimental validation. This approach aligns well with tasks that demand both innovation and deployment readiness, such as automated PII detection in EHRs, and has been reaffirmed in recent data science literature emphasizing methodological thoroughness and real-world applicability (Leunget al., 2023; Jahani et al., 2023). In this dissertation, constructive research guides the design, implementation, and evaluation of an ML pipeline capable of detecting direct and indirect identifiers in unstructured clinical text. The process follows iterative cycles that incorporate prior research insights while adapting to scalability, adaptability, and regulatory compliance requirements.

Constructive design over alternatives such as experimental or case study approaches. While experimental research isolates and tests specific hypotheses under controlled conditions, it lacks the flexibility to support full-cycle system development. Although helpful in generating context-specific insights, case studies do not facilitate the iterative construction and refinement of generalizable tools. In contrast, constructive research supports the systematic development of deployable systems tailored to real-world challenges while maintaining methodological integrity and cross-context applicability (Jahani et al., 2023).

Despite its advantages, constructive research presents limitations. It may overemphasize technical artifact development at the expense of contextual interpretation, and findings may lack generalizability if the evaluation remains narrowly scoped. In addressing these concerns, this study applies validation across diverse datasets (e.g., MIMIC-III), incorporates feedback loops throughout development, and aligns the methodology with regulatory and ethical frameworks such as HIPAA and GDPR. These strategies ensure that the resulting system performs effectively and remains legally and ethically compliant (Leung et al., 2023).

The artifact developed in this study is an NLP-driven ML pipeline that combines transformer-based language models and unsupervised clustering algorithms to detect both direct identifiers (e.g., names, Social Security numbers) and indirect identifiers (e.g., rare conditions, occupations) in unstructured clinical text. Constructive methodology was selected because it emphasizes iterative artifact construction, feedback-informed refinement, and alignment with empirical testing workflows common in ML. Unlike case studies or purely observational research designs, constructive methodology emphasizes building and validating, allowing this dissertation to contribute a functioning tool and generalizable insights into scalable PII detection.

A standard limitation of constructive research is its potential to yield context-dependent solutions that may not generalize across domains. To address this, the methodology in this study includes model validation across heterogeneous EHR datasets—including structured and unstructured components of MIMIC-III—and incorporates fairness-aware design strategies to enhance generalizability and demographic robustness. The framework is stress-tested using multiple evaluation metrics such as F1-score, precision, recall, silhouette score, and Davies-Bouldin index, ensuring technical soundness and equity.

Researchers and practitioners use the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework (Ma et al., 2025) as a structured, iterative approach to developing data-driven systems across six phases: business understanding, data understanding, data preparation, modeling, evaluation, and deployment. Health informatics and privacy-preserving data mining applications, such as PII detection, apply the CRISP-DM framework to structure ML pipelines systematically. For example, extensions such as CRISP-ML(Q) enhance the original framework by embedding legal compliance (e.g., with HIPAA and GDPR) and ethical quality assurance into each stage (Saltz et al., 2023), reinforcing its relevance in privacy-focused healthcare applications. Although CRISP-DM remains widely influential, critics note its limited engagement with ethical and quality assurance concerns. CRISP-ML(Q) extends the CRISP-DM framework by embedding quality assurance, ethical considerations, and regulatory compliance, such as HIPAA and GDPR, into the ML lifecycle, making it well-suited for healthcare applications (Sankaran et al., 2023; Hirt et al., 2020; Studer et al., 2020). The CRISP-DM framework aligns with constructive design methodology by supporting iterative, artifact-centered development across six structured phases, facilitating reproducible, privacy-preserving ML pipelines tailored to sensitive clinical data environments.

Unlike traditional approaches that rely on static rule-based systems, ML models exhibit dynamic learning capabilities and improve accuracy by assimilating patterns from diverse data sources (Ambati, 2025). The proposed framework in this research incorporates NLP techniques for entity recognition and classification while utilizing unsupervised learning strategies, such as clustering to structure data while using a constructive development approach, iterating through various model architectures and tuning methodologies to enhance privacy-preserving capabilities through recent advancements in healthcare data privacy (Ancker et al., 2024).

The presence and accurate identification of PII remains a critical factor influencing the reliability of data sharing and the ethical soundness of ML models designed to detect PII within EHRs. With the growing adoption of automated de-identification techniques in response to evolving privacy regulations—including HIPAA, the GDPR, and the EU AI Act—there is increasing reliance on automated PII detection to achieve both model effectiveness and regulatory compliance. This reliance necessitates access to high-quality, well-annotated datasets. Three intersecting dimensions of EHR data quality—incompleteness, inaccuracy, and institutional heterogeneity—consistently degrade the effectiveness and fairness of ML models in this domain.

Incompleteness arises from missing values, sparse demographic attributes, or absent clinical notes. These omissions frequently result from non-standardized documentation workflows across healthcare providers. Sarkar et al. (2024) report that datasets with over 15% missing fields can reduce the accuracy of transformer-based PII detection models by up to 37% and elevate false-negative rates by 41%. Wu et al. (2022) similarly confirms that absent fields disrupt feature representation and embedding layers, especially in models reliant on positional or contextual tokenization.

Inaccuracy, including mislabeled tokens, misspelled identifiers, and corrupted timestamps, injects semantic noise disproportionately affecting model precision. For instance, Morafah (2025) found that phonetic name variants (e.g., "Jon" vs. "John") resulted in a 31% rise in false-positive rates within FL environments; these errors undermine the trustworthiness of outputs and increase the likelihood of data retention violations when systems incorrectly preserve PII. Wu et al.(2022) emphasized that algorithmic complexity alone could not correct error-prone EHR data; BERT-based models demonstrated a 48% variance in recall across datasets differing solely in quality auditing status.

Institutional heterogeneity, a defining feature of FL and multi-site model training, presents a formidable barrier to consistency. While FL promises enhanced privacy through decentralized data training, Morafah (2025) emphasize that differing data entry and annotation standards among participant institutions distort gradient aggregation, leading to model drift and inequitable performance. Yang et al. (2023) further illustrates the impact of data source bias by showing how the underrepresentation of rural clinical data reduced generalizability and fairness, with F1-score drops of 0.18 in cross-institution validation; these findings reinforce the need to couple technical solutions with systematic data quality interventions by embedding mitigation strategies—such as automated imputation, noise detection modules, and normalization pipelines—within the model development lifecycle. Sarkar et al. (2024) and Yang et al. (2023) advocate for continuous quality audits to enhance model accuracy while supporting regulatory compliance and ethical alignment.

Morafah (2025) further proposes integrating real-time validation layers into FL aggregation mechanisms to monitor model drift and institutional bias in distributed systems.

This work collectively emphasizes that algorithmic sophistication cannot overcome foundational data flaws. As regulatory scrutiny of health AI increases in 2025, ML systems that fail to address these persistent quality issues risk violating legal standards and undermining patient trust. Consequently, ensuring data completeness, accuracy, and cross-institutional harmonization is not merely a best practice—it is a foundational requirement for the ethical deployment of PII detection technologies in health informatics. Erroneous data, such as misspelled names, incorrect dates, or inconsistent formatting, disrupts the contextual patterns models like BERT rely on for accurate entity recognition, leading to unreliable detection outcomes (Wu et al., 2022). Furthermore, biased data—resulting from underrepresentation of certain demographic groups or skewed clinical documentation—can cause ML models to misclassify PII in minority patient records disproportionately, exacerbating healthcare disparities (Yu et al., 2025). Addressing these challenges requires robust data pre-processing techniques, such as imputation for missing values, standardization to correct errors, and bias mitigation strategies, to ensure ML models achieve reliable PII detection while adhering to privacy regulations like HIPAA and GDPR

By integrating ML models with privacy-enhancing architectures, specifically FL and differential privacy, this study proposes a scalable and adaptable framework for secure healthcare data analysis. Federated learning decentralizes ML by training models directly on multiple devices or institutional servers without transferring raw data to a central repository. This architecture mitigates the risk of direct data exposure by ensuring PHI remains locally stored, a crucial consideration for HIPAA compliance and institutional data governance standards (Kairouz et al., 2021). Differential privacy, by contrast, is a statistical technique that introduces mathematically calibrated noise into datasets or model outputs to prevent the re-identification of individuals within a dataset, even in the presence of auxiliary information (Im et al., 2024). When embedded into FL pipelines, differential privacy further enhances confidentiality by obfuscating model updates or aggregated outputs without substantially degrading model utility.

FL and differential privacy enable the collaborative use of sensitive health data across institutions while preserving individual privacy and satisfying legal obligations under regulations such as HIPAA and GDPR. This study employs these technologies to iteratively develop, test, and validate a privacy-preserving ML framework that supports ethical data reuse in clinical research and operational environments. Through such integration, the framework balances data accessibility and privacy protection—key concerns in real-world health informatics deployments (Subramanian et al., 2024).

This dissertation proposes a contribution to advancing ML-driven privacy solutions by employing a constructive design approach that enables the development of a systematic and adaptable methodology for PII detection in large-scale EHR datasets. The iterative framework design supports accuracy, efficiency, and practical deployment within privacy-sensitive environments.

Identifying challenges and opportunities related to data availability and PII concerns is essential. Appiah et al. (2021) argue that innovative governance models, including ML, must balance the pressures of maintaining privacy and maximizing economic outcomes. This governance approach can unlock the value of data while adhering to ethical and legal standards. Frameworks such as the OECD AI Principles and the AI RMF (NIST, 2023) provide structured guidelines for governance, transparency, and accountability in AI systems, addressing the evolving intersection of ML and regulatory compliance in healthcare. Wei et al. (2022) introduced an ML framework that enhanced PII detection by targeting quasi-identifiers and contextual metadata elements frequently overlooked by rule-based or pattern-matching algorithms. The framework applied feature extraction techniques that quantified attribute interdependencies and improved the identification of latent privacy risks in EHR data. Ahmed et al. (2020) emphasized AI-based anonymization techniques, including differential privacy and homomorphic encryption (a method that allows data to stay encrypted even while being analyzed), to mitigate re-identification risk while maintaining data utility. Both studies addressed balancing individual privacy and analytical validity, particularly when applying automated methods to large-scale, high-dimensional public health datasets.

Silva et al. (2022) introduce a privacy risk assessment framework that combines ML and traditional techniques, focusing on minimizing privacy risks while maintaining system performance. The challenge of acquiring high-quality training data, especially in open public datasets, is similarly discussed by Wu et al. (2022), who stresses the risks of PII exposure online and the need for ML tools to safeguard sensitive data.

While this literature provides valuable insights into balancing data accessibility and privacy, this challenge is especially evident in large-scale public datasets. Current models often need help with this balance, as seen in public administration, where data sharing is widespread. Yukhno (2024) points out a need for more sophisticated governance structures integrating ML to manage these risks.

This research explores facilitating data access while mitigating data exposure risks. This approach could guide the development of ML models focused on minimizing risk through PII detection while ensuring data accessibility. Differential privacy, as discussed by Xu et al. (2022), offers a structured method to balance utility and privacy by adding controlled noise to sensitive data. Recent advancements in computational methods have enabled educational and archival institutions to implement tools for managing PII in archival records. These methods aim to improve accuracy, transparency, and privacy compliance, forming the basis for scalable and reproducible PII protection systems in data-intensive domains (Canonici & Caruso, 2025). It also demonstrates how institutions can balance privacy with data accessibility to serve the public interest and how regulatory pressures necessitate more advanced PII detection solutions.

Delgado Ramos (2023) discussed computational methods for handling sensitive data, including Optical Character Recognition (OCR), a technique that converts scanned images of text into machine-readable text, and NER, an NLP technique that automatically identifies and classifies entities such as names, dates, and locations in unstructured text. When combined, OCR enables the digitization of archival records. At the same time, NER facilitates the automated detection and redaction of PII within those records. These integrated tools support PII protection and promote data accessibility by balancing individual privacy with the public's interest in historical data, laying the groundwork for transparent data governance in sensitive archival contexts emphasized in this study. This foundation supports exploring ML models that can learn and adapt over time, enhancing the effectiveness in identifying and protecting PII in EHR. Such models align with the purpose of the study and their research questions, focusing on how ML models can detect PII without compromising data availability.

The history of digitization of healthcare records began in the 1960s with the advent of EHRs, which aimed to improve data accessibility and patient care (Zhang et al., 2024). However, the rise of EHRs also introduced significant privacy concerns, particularly regarding protecting PII. Early efforts to address these concerns focused on regulatory frameworks, such as the HIPAA in 1996, which established standards for safeguarding sensitive patient data (Numajiri & Hayashi, 2024). Despite these regulations, the increasing volume of data generated by EHRs has made manual PII detection and protection increasingly challenging, necessitating the development of automated solutions.

History

The evolution of this research topic reflects the growing complexity of balancing data accessibility and privacy in healthcare. Initially, privacy protection relied heavily on policy and manual oversight. However, the exponential growth of healthcare data and the advent of ML in the 2000s shifted focus toward automated solutions. Researchers began exploring ML techniques to identify and protect PII in EHRs, leveraging advancements in NLP and pattern recognition (Kovačević et al., 2024). This shift marked a significant milestone in the field, as ML-based approaches offered scalable and efficient alternatives to traditional methods.

Key milestones and development of HIPAA established the foundation for protecting patient data, emphasizing the need for secure handling of PII in EHRs (Numajiri & Hayashi, 2024). This regulation highlighted the tension between data accessibility and privacy, setting the stage for future research. Applying ML techniques to healthcare data marked a turning point in PII detection. Early studies demonstrated the potential of ML algorithms to identify sensitive information in unstructured EHRs (Kovačević et al., 2024). Advances in NLP enabled more accurate identification of PII in free-text EHRs. Techniques such as NER and deep learning models improved the precision and recall of PII detection systems (Durango et al. 2023). Recent research has focused on integrating privacy-preserving techniques into ML models, such as FL and differential privacy. These approaches aim to enhance data accessibility while minimizing privacy risks (Kaissis et al., 2020). The ongoing development of ML-based PII detection systems has prompted discussions about ethical and regulatory implications. Researchers and policymakers are working to establish guidelines for the responsible use of ML in healthcare data management (Al Zaabi, & Alhashmi, 2024).

The framework for balancing data accessibility and privacy in EHRs using ML PII detection has evolved from the intersection of data privacy regulations, advancements in ML algorithms, and the growing need for secure data sharing in healthcare. The HIPAA in the U.S. and the GDPR in the EU have been pivotal in shaping the need for robust PII detection mechanisms (Shojaei et al., 2024). Early approaches relied on rule-based systems for PII detection, but these were limited in scalability and adaptability. The advent of deep learning has enabled more sophisticated and accurate detection of PII in unstructured EHR data (Ozonze et al., 2023).

Recent advancements in NLP have led to the development of transformer-based models that have enhanced the detection of PII in text data. For instance, the fine-tuned GPT-4o-mini model has demonstrated superior performance in PII detection tasks, achieving a recall of 0.9589 on the CRAPII dataset (a benchmark dataset used to test how well models can detect personal information in text), surpassing traditional frameworks like Microsoft Presidio and Azure AI-Language in both accuracy and computational efficiency (Shen et al., 2025). Pre-trained on large corpora and fine-tuned for specific tasks, transformer-based models demonstrate high effectiveness for PII detection in EHRs. Additionally, researchers have adopted FL as a promising approach to train ML models on decentralized data without compromising privacy (Chen et al., 2020); this approach aligns with the need to balance data accessibility and privacy by enabling collaborative model training while keeping data localized. Several studies have employed ML frameworks for PII detection in EHRs. Liu et al. (2020) combined convolutional neural networks (CNNs) and recurrent neural networks (RNNs) to detect PII in clinical notes, reporting over 94% accuracy on i2b2 datasets (a publicly available set of actual hospital notes used to train and test medical AI models). While effective at the time, this approach lacked contextual depth and struggled with long-range dependencies. Subsequent developments, such as those by (Han et al., 2022), introduced BERT-based models that outperformed earlier CNN-RNN architectures, achieving higher precision and recall by leveraging contextual embeddings and attention mechanisms. These transformer-based models have since become the foundation for state-of-the-art de-identification systems. Complementary privacy-preserving frameworks have also progressed. Recent implementations of differential privacy incorporate adaptive noise-injection techniques to preserve utility in sensitive health datasets (Vidanage et al., 2022), while advances in homomorphic encryption now support more efficient encrypted model inference with reduced computational overhead (Tertulino et al., 2024). These innovations collectively reflect a shift toward integrated solutions that balance privacy, accuracy, and scalability in real-world clinical NLP applications. However, these methods often compromise data utility, making it less suitable for applications requiring high data accessibility.

The selected framework integrates transformer-based ML models, such as BERT, with FL. This decentralized training approach avoids centralized data aggregation. BERT enables advanced contextual understanding of unstructured clinical text, improving the precision of PII detection across complex narrative structures. Federated learning facilitates model training across distributed healthcare institutions without requiring raw data transfers, thereby reducing privacy risks while maintaining model accuracy; a combination that addresses the dual challenge of data accessibility and privacy preservation in healthcare environments. The framework promotes compliance with regulatory standards such as HIPAA by incorporating privacy-preserving architecture directly into the training pipeline. It supports scalable deployment in real-world clinical research settings. Unlike differential privacy and homomorphic encryption, this framework maintains high data utility while ensuring robust privacy protection. It also aligns with the CRISP-DM methodology, which provides a structured approach to developing ML models (Ma et al., 2025).

The development of algorithms for PII detection in EHRs is grounded in mathematical and probabilistic theories, such as probability distributions used to model uncertainty in text classification. For instance, transformer models like BERT use attention mechanisms grounded in linear algebra and probability theory (Huang et al., 2024). These mechanisms allow the model to weigh the importance of different words in a sentence, making it highly effective for PII detection. Additionally, FL employs optimization techniques such as stochastic gradient descent (SGD) to train models on decentralized data (Dang et al., 2022).

The CRISP-DM methodology guides the design and implementation of this dissertation's ML framework by providing a structured process for data analysis. This research replicates the six phases of the CRISP-DM procedure: business understanding, data understanding, data preparation, modeling, evaluation, and deployment (Ma et al., 2025); this methodology ensures that the development of ML models for PII detection remains systematic and aligned with the study's objectives.

The selected framework is relevant to the present study as it addresses the core challenge of balancing data accessibility and privacy in EHRs. ML models like BERT ensure accurate PII detection, while FL enables secure model training on decentralized data. This framework guided the development of the problem statement, which focuses on the need for robust PII detection mechanisms in EHRs, and the purpose statement, which aims to develop an ML-based solution that balances data accessibility and privacy. The research design explores the effectiveness of this framework in real-world healthcare settings, ensuring both theoretical grounding and practical relevance.

Data Ethics and Legal Frameworks

Ensuring data ethics represents the foundational rationale for this research study, informing the design method, deployment, and governance of ML systems or models for detecting personal PII in EHRs. As digital infrastructures increasingly rely on data-driven decision-making, ethical and legal frameworks are essential to safeguard individual rights, promote transparency, and ensure accountability. This study emphasizes and promotes that technical innovation in health informatics must be grounded in principled, legally compliant, and socially responsive data practices.

Data ethics applies practical moral principles to data collection, processing, analysis, and use. Classical ethical theories provide a framework for interpreting dilemmas arising from data science applications. Utilitarianism, for instance, demands the maximization of collective benefit (such as data-driven decisions and PII protection) and is often invoked when balancing public health research advantages against potential privacy harms (Hanna et al., 2025). Deontological ethics (which emphasizes adherence to moral duties and principles regardless of consequences, such as respecting individual rights to consent, privacy, and autonomy) mandates adherence to moral duties, including informed consent, data subject rights, and non-maleficence, regardless of the outcome. Virtue ethics, a normative ethical theory that emphasizes moral character and the cultivation of professional integrity over rule-based or outcome-based reasoning, centers on character and demands that data scientists uphold honesty, fairness, and diligence in serving humanity (Hagendorff, 2022).

These ethical orientations converge on core principles essential to contemporary data governance: privacy, fairness, transparency, autonomy, and accountability. Adhering to these principles is a moral imperative and a practical necessity for building public trust in AI systems and data infrastructures (Maciejewski et al., 2025).

Global data protection laws establish binding obligations on entities that process personal data. The European Union enacted the GDPR in 2018, establishing it as the most influential framework by codifying key principles such as data minimization, purpose limitation, transparency, and the right to be forgotten (Conduah et al., 2025). GDPR's extraterritorial applicability obligates international compliance and has prompted the adoption of comparable frameworks in other jurisdictions.

In the United States, the HIPAA governs health information privacy, mandating the protection of 18 classes of protected health information (PHI). HIPAA emphasizes de-identification, consent management, and breach notification protocols. Complementary regulations, including the California Consumer Privacy Act (CCPA) and the 21st Century Cures Act, expand data portability and digital access rights.

Policymakers are increasingly aligning emerging regulatory developments with AI oversight. The proposed European Union Artificial Intelligence Act (EU AI Act) classifies AI applications by risk level. It imposes transparency, human oversight, and documentation standards on high-risk systems, including those processing health-related data. Similarly, the U.S. National Institute of Standards and Technology (NIST) released the AI Risk Management Framework (2023), which guides developers in ensuring fairness, security, and privacy by design in AI systems.

The Fair Information Practices (FIPs), initially articulated by the U.S. Federal Trade Commission, serve as a global normative foundation. The principles include notice, choice, access, security, and enforcement, embedded within legal systems and professional guidelines. Adhering to FIPs ensures that data processing respects individual autonomy and institutional accountability (Ambati, 2025).

Simultaneously, open data initiatives such as the U.S. Open Government Data Act and the international Open Data Charter promote accessibility and reuse of public datasets. These efforts drive transparency and innovation but introduce risks related to re-identification, algorithmic bias, and surveillance. Addressing these tensions requires robust ethical safeguards. The FAIR data principles (Findable, Accessible, Interoperable, Reusable) complement open data strategies by emphasizing stewardship obligations, especially in biomedical contexts (Vogt et al., 2024).

The proliferation of LLMs such as GPT-4 and domain-specific transformers in health applications introduces new challenges for data ethics. These models often require massive datasets, some of which may contain sensitive or insufficiently anonymized health records. Recent studies show that LLMs can inadvertently memorize and regurgitate identifiable information from training corpora, raising risks of privacy breaches (Carlini et al., 2023). Deployment in clinical settings demands rigorous de-identification, continuous monitoring, and adversarial testing.

Generative AI further complicates traditional consent models by enabling inference from synthetic data or large-scale embeddings that may reveal sensitive traits, even when systems remove direct identifiers. To address this, privacy-preserving ML (PPML) techniques—including differential privacy, homomorphic encryption, and federated learning—are essential for mitigating re-identification risks while maintaining data utility (Guerra-Manzanares et al., 2023).

Federated learning offers structural advantages for healthcare: models are trained locally on edge devices or institutional servers, thus reducing centralization risks. This decentralized approach aligns with HIPAA and GDPR requirements, supporting security and accountability.

The ethical governance of AI and data systems is rapidly evolving. Anticipated developments include the expansion of AI-specific regulatory regimes (e.g., EU AI Act), global convergence of privacy standards (e.g., OECD AI Principles), and stronger mandates for algorithmic accountability. Additionally, there is growing recognition of the need to address data justice, which focuses on structural inequities in data collection and usage practices (Pierce & Cleary, 2024).

The future of data ethics also lies in embedding or incorporating ethical reasoning directly into AI systems through explainable AI (XAI), value-sensitive design, and continuous suitability. The IEEE's "Ethically Aligned Design" initiative (2021) and UNESCO's 2022 Recommendation on the Ethics of Artificial Intelligence offer pathways for embedding ethical and human rights considerations in digital infrastructure.

As ML models increasingly mediate access to healthcare services, employment, and public benefits, ensuring these technologies reflect ethical, inclusive, and lawful practices is imperative. This study contributes to that effort by designing a privacy-preserving ML framework that foregrounds ethical compliance, aligns with HIPAA/GDPR standards, and anticipates the challenges introduced by generative AI and LLMs.

This study's ML framework for detecting PII in EHRs is grounded in a comprehensive understanding of data ethics, informed by classical ethical theories, evolving legal mandates, and emerging governance standards in AI through the integration of privacy-preserving techniques—such as differential privacy and FL—the framework ensures compliance with the GDPR and the HIPAA while addressing the specific privacy risks posed by LLMs and other AI-driven architectures.

This research framework exemplifies responsible data handling in health informatics by incorporating ethical reasoning about PII into technical design and aligning with global regulatory expectations. Anticipated developments, including AI-specific legislative frameworks, harmonized international privacy standards, and embedded auditing mechanisms, are expected to strengthen the applicability and ethical resilience of the proposed framework, fostering public trust and enabling secure, equitable, and lawful data-driven innovation in healthcare.

Traditional Methods of PII Detection in EHRs

Traditional approaches to PII detection in EHRs, such as manual review, rule-based systems, and basic anonymization techniques, have been widely employed but face limitations (Vahdati et al., 2024). Manual review, while thorough, is impractical for large-scale datasets due to its labor-intensive nature. Rule-based detection relies on predefined patterns (e.g., Social Security numbers or email formats) but struggles with unstructured data and evolving PII formats—recent research critiques rule-based methods for failing to adapt to the complexity of modern EHR systems. For example, Hong et al. (2024) emphasizes the limitations of rule-based systems, which often overlook contextual PII, such as names embedded in free-text notes, resulting in incomplete de-identification. Regulatory frameworks like HIPAA and GDPR have shaped traditional methods by mandating strict compliance; however, conventional methods often lag behind technological advancements (Xu et al., 2022). Im et al. (2024) argue that traditional anonymization, such as masking or pseudonymization, compromises data utility, hindering downstream analytics, a key concern for ML-driven alternatives.

Recent advancements in PII detection have focused on integrating rule-based and AI-driven methodologies to enhance precision and scalability. Traditional methods, such as deterministic regular expressions (DREs; structured pattern-matching rules that operate with predictable and unambiguous outcomes) and manually curated entity recognition rules, often fall short when applied to large-scale or heterogeneous EHR datasets due to limited adaptability and high false favorable rates. To address these limitations, hybrid frameworks have emerged that combine rule-based systems with supervised ML models, including support vector machines (SVMs; algorithms that classify data by finding the optimal boundary that separates categories), decision trees, and neural networks—resulting in improved contextual understanding and classification accuracy (Negash et al., 2023).

Researchers have employed NLP techniques powered by pre-trained language models like BERT to detect nuanced, indirect, or context-dependent PII in clinical narratives. These models improve semantic comprehension (i.e., the ability to understand contextual meaning and relationships between terms within the text) and support fine-tuned detection of domain-specific features, thereby reducing errors in identifying sensitive data.

Recent research also explores lightweight transformer models, such as DistilBERT and TinyBERT, trained using privacy-preserving techniques like differential privacy and adversarial regularization to support real-time, on-device PII detection without sacrificing accuracy or violating data protection constraints (Kuo et al., 2025)

ML Approaches for PII Detection

ML has become a pivotal methodology in detecting PII, particularly in managing the scale and heterogeneity of EHR data. Both supervised and unsupervised learning techniques contribute to this effort by enabling systems to identify patterns in structured and unstructured clinical narratives (Altalla’ et al., 2025). Transformer-based models, such as BERT, demonstrate strong performance in detecting PII within free-text fields by leveraging contextual embeddings to discern term meaning based on linguistic environment (Chen et al., 2020). Recent refinements in

BERT architecture, including domain-specific fine-tuning, have yielded detection accuracy exceeding 95 percent across various EHR formats, outperforming traditional rule-based systems (Altalla’ et al., 2025). However, these gains come at the cost of increased computational demands, which may limit feasibility in real-time or low-resource settings (Chen et al., 2020).

Researchers have adopted federated learning (FL) to address privacy and scalability concerns. This decentralized training paradigm allows local model updates without transmitting raw data. FL preserves data confidentiality while maintaining model performance; however, it also presents challenges such as communication latency, model drift, and complex optimization requirements (Liu et al., 2023; Kuo et al., 2025). Complementing these approaches, Kohli et al., (2025) propose hybrid models that integrate reinforcement learning with transformer architectures, offering enhanced scalability and reduced resource consumption. In parallel, researchers are integrating explainable AI (XAI) frameworks with PII detection models to improve interpretability and transparency, particularly critical in high-stakes domains like healthcare, where clinicians and auditors must understand how sensitive information is flagged or omitted. XAI techniques, such as attention visualization and feature attribution, provide insights into model decisions and support regulatory compliance with accountability standards (Huang et al., 2024). Collectively, these advancements illustrate the increasing superiority of ML-driven frameworks over traditional techniques in terms of accuracy, efficiency, interpretability, and alignment with evolving privacy-preserving standards.

Transformer-based NER models, such as BERT, perform well in detecting closed-set PII, which refers to predefined and structured entity types, including names, dates, addresses, and hospital IDs. These identifiers follow predictable linguistic and syntactic patterns that enable supervised models to learn clear decision boundaries using annotated training data (Li et al., 2024).

However, EHRs often contain open-set PII, which consists of context-dependent, rare, and irregular identifiers such as unique disease mentions, family relational cues (e.g., “my brother Ahmed”), occupational identifiers (e.g., “works at Valley Clinic”), or inferred identities based on location or social context (Han et al., 2023). These irregular identifiers lack stable labeling patterns and cannot be captured reliably using supervised BERT-based detection alone (Chung et al., 2023). In these cases, unsupervised clustering models such as DBSCAN are effective because it detects anomalous density regions in embedding space, identifying context-sensitive or rare identifiers that fall outside predefined entity classes (Chung et al., 2023). Integrating DBSCAN with transformer embeddings, therefore, enables detection of both closed-set and open-set PII, achieving a broader and context-aware privacy protection mechanism that aligns with Research RQ 2 and RQ 5.

For domain-specific applications, specialized transformer-based models such as BioBERT, ClinicalBERT, and BlueBERT are pre-trained on biomedical and clinical text corpora, including PubMed abstracts, full-text articles, and the MIMIC-III critical care database. Transformer-based models provide improved contextual embeddings for medical terminology, enhancing performance in de-identification and clinical NER tasks, including the extraction of diagnoses, medications, and personal identifiers (Lee et al., 2020; Li et al., 2023; Lee et al., 2020). In settings with limited computational resources, non-transformer architectures such as ELMo and FastText, when combined with Conditional Random Fields (CRFs), offer viable albeit lower-performing alternatives by generating contextualized or subword-based embeddings that support baseline PII detection in EHRs (Clark et al., 2020). Additionally, instruction-tuned models—such as T5, FLAN-T5, and GPT-NeoX—facilitate few-shot learning and prompt-based sequence labeling, allowing rapid adaptation to evolving annotation schemas or emerging de-identification tasks with minimal labeled data. These alternatives present promising avenues for empirical evaluation, particularly in real-world deployments with domain adaptation constraints or fluctuating annotation requirements.

Future directions in privacy-preserving ML (PPML) emphasize advancing confidentiality through differential privacy, homomorphic encryption, and secure multiparty computation. These techniques mitigate disclosure risks while enabling collaborative model training across distributed healthcare systems. Concurrently, researchers are expanding quantum ML (QML) to reduce computational burdens associated with large-scale transformer models while maintaining detection accuracy (Corli et al., 2024). Researchers and developers are creating ethical auditing mechanisms and integrating explainable AI (XAI) tools to enhance transparency, foster stakeholder trust, and ensure alignment with emerging regulatory standards such as the EU AI Act and HIPAA (Huang et al., 2024). These innovations collectively aim to reconcile technical efficacy with legal and ethical imperatives in automated PII detection systems. Therefore, this research employs a hybrid framework in which transformer-based contextual embeddings extract semantic representations, and DBSCAN isolates anomaly-driven PII exposure zones beyond traditional entity classes.

NLP and NER for PII Identification

NLP techniques, particularly NER, have revolutionized PII detection by extracting sensitive entities from unstructured EHR text (Delgado Ramos, 2023). Traditional NER models, such as Conditional Random Fields (CRF), have evolved into deep learning–based approaches that enhance precision through contextual understanding (Durango et al., 2023). Recent studies showcase significant progress in this domain. For instance, Li et al. (2023) introduce a BiLSTM-CRF model—where BiLSTM stands for Bidirectional Long Short-Term Memory, a recurrent neural network architecture capable of capturing both past and future context in a sequence, and

CRF refers to a probabilistic model used for structured prediction. Integrating BiLSTM with CRF enables the model to leverage temporal context and label dependencies simultaneously, resulting in improved sequence tagging performance. When enhanced with attention mechanisms, this architecture achieved a 10% reduction in false positives compared to earlier methods. In addition to traditional and deep learning models, researchers have increasingly explored large language models (LLMs)—such as GPT-4 (Generative Pre-trained Transformer4)—for zero-shot and few-shot NER tasks. These transformer-based models, pre-trained on massive text corpora, demonstrate the capacity to generalize across domains and perform entity recognition without extensive retraining (Chang & Bergen, 2024). GPT-4 offers advanced contextual reasoning and semantic comprehension, enabling it to detect PII with minimal task-specific supervision. This capability makes LLMs especially valuable in scenarios involving diverse and evolving clinical documentation styles. These advancements emphasize the important role of NLP in improving both accuracy and adaptability while addressing the shortcomings of rule-based systems in handling free-text EHR data (Chang & Bergen, 2024).

Recent advancements in NLP have expanded to include multilingual models for PII detection, proving effective in healthcare environments where clinical documentation is not limited to English (Garcia et al., 2024). Researchers have also begun integrating NLP with FL to preserve privacy while maintaining high detection accuracy in decentralized EHR systems (Liu & Zhang, 2024). These FL-based frameworks support distributed model training without transmitting raw data, maintaining compliance with HIPAA and GDPR while enabling collaborative analytics across institutions. Recent studies demonstrate that coupling FL with anonymization strategies enhances secure and scalable PII detection while mitigating the risk of data leakage (Dang et al., 2022).

Privacy-Preserving Techniques in ML

Mathematical privacy, in contrast, centers on the use of formal guarantees to minimize the risk of individual-level data disclosure in statistical analyses; differential privacy, the most established form of mathematical privacy, was formalized by Im et al. (2024) to introduce calibrated noise to data outputs, thereby ensuring that the presence or absence of any single individual does not significantly affect analytical results. The privacy budget parameter epsilon (ε) quantifies the strength of this guarantee, with smaller values indicating stronger privacy. Differential privacy is particularly relevant in ML-based PII detection because it allows high-utility model training while rigorously bounding the risk of re-identification. Alsaigh et al. (2024) demonstrate that deep learning tasks in healthcare benefit from differential privacy due to the pronounced sensitivity of individual data points. Moreover, Nava et al. (2023) propose a transformer-based ML architecture incorporating differentially private gradient perturbation, achieving F1-scores exceeding 0.85 while maintaining $\varepsilon < 1$ across large-scale EHR corpora. The work demonstrates the viability of privacy-compliant deep learning pipelines, aligning with the study's objectives.

The practical significance of these methods lies in protecting patient confidentiality while enabling meaningful data analysis—a dual imperative in contemporary health informatics. According to the U.S. Department of Health and Human Services (Shojaei et al., 2024), over 109 million individuals were affected by healthcare data breaches in 2023, underscoring the urgency of adopting provable privacy measures. Homomorphic encryption prevents raw data exposure during processing, mitigating risk during ML operations. Meanwhile, differential privacy ensures statistical resilience against inference attacks, essential when working with sensitive unstructured data from sources such as the MIMIC-III dataset. These techniques collectively support regulatory compliance with HIPAA, GDPR, and the evolving EU AI Act, reinforcing methodological design's ethical and legal validity.

Emerging literature further illustrates the relevance of combining these approaches within FL frameworks—an architecture central to this dissertation. FL enables decentralized model training across institutional nodes without direct data exchange, reducing breach risk while accommodating data sovereignty constraints. Ahmed et al. (2020) demonstrate that integrating FL with differential privacy and homomorphic encryption can significantly enhance privacy guarantees, particularly in heterogeneous data environments daily in healthcare. Jiang et al. (2022) reports a 30% reduction in multi-institutional breach risks when adopting hybrid methods. Additionally, Tölle et al., 2025) emphasize that these technologies democratize access to secure data analytics for under-resourced institutions, supporting dissertation's focus on scalable, equitable ML infrastructures. Recent studies have further explored using secure multi-party computation (SMPC) in conjunction with differential privacy to enhance the security of collaborative ML pipelines without compromising data utility (Taylor et al., 2024). Additionally, researchers have proposed integrating blockchain with privacy-preserving ML methods to establish tamper-proof logging mechanisms for PII detection workflows, which enhance audibility and traceability in compliance with HIPAA and GDPR standards (Brown et al., 2024). Kovačević et al. (2024) further strengthen ethical considerations by arguing that mathematical privacy reduces bias and re-identification risk and improves model transparency—an essential criterion for fairness-aware system design.

Using Microsoft SEAL for homomorphic encryption and TensorFlow Privacy for differential privacy in a BERT-based NLP pipeline, guided by the CRISP-DM methodology, provides a practical foundation for privacy-preserving ML in healthcare (Han et al., 2022).

Alsaigh et al. (2024) introduce a hybrid homomorphic encryption scheme optimized for EHRs, allowing secure computation on encrypted PII datasets. These techniques comply with regulations like GDPR and enhance data utility for research, addressing the accessibility-privacy trade-off (Kaissis et al., 2020). Ensuring the suitability and provenance of data within ML pipelines is foundational for building trustworthy, transparent, and accountable health informatics systems. These attributes refer to the capacity of an ML system to trace data origin (provenance) and assess whether the data and processing logic are appropriate for a given predictive task (suitability). Provenance supports reproducibility, auditability, and regulatory compliance—key requirements in domains governed by data protection regulations such as HIPAA and GDPR. Suitability, in turn, ensures that training data aligns with the intended context of model use, mitigating risks of dataset shift and model failure in deployment (Ahmed et al., 2023; Gebru et al., 2021).

Researchers have developed structured documentation approaches such as datasheets for datasets (Gebru et al., 2021) and model cards (Liu et al., 2024) to formalize these concepts in ML practice. n healthcare, practitioners increasingly adopt these documentation practices to ensure traceability, particularly when clinical decisions rely on algorithmic outputs. Ahmed et al. (2023) further demonstrate that incorporating data lineage tracking into ML workflows enhances stakeholder trust and supports external audits. These findings validate the importance of data provenance and suitability as part of fairness-aware and ethically aligned ML pipeline design—an approach central to this dissertation's proposed framework.

Ethical and Legal Considerations in PII Detection

The deployment of ML for PII detection raises ethical and legal challenges, including algorithmic bias, data ownership, and compliance with regulations like GDPR, CCPA, and HIPAA (Maciejewski et al., 2025; Conduah et al., 2025). Recent scholars emphasize these issues. Yu et al., (2025) argue that biased ML models can disproportionately misclassify PII in minority patient groups, exacerbating healthcare disparities. Legally, Brown et al. (2024) note that evolving regulations, such as the EU AI Act, impose stricter accountability on ML systems, requiring transparency in PII processing. Balancing open data initiatives with privacy remains contentious, as Subramanian et al. (2024) highlighted, underscoring the need for ethical frameworks to guide ML deployment.

Challenges in Implementing ML-Based PII Detection

Implementing ML for PII detection in EHRs requires addressing several interrelated challenges, including data heterogeneity, regulatory constraints, and model interpretability (Wu et al., 2022). This research adopts a hybrid ML approach that integrates a fine-tuned BERT model with unsupervised clustering techniques—specifically, density-based spatial clustering of applications with noise (DBSCAN)—and support vector machines (SVMs) for classification. This approach balances semantic contextualization, structural inconsistency across data formats, and classification robustness in diverse healthcare environments.

Recent literature emphasizes various ML methods for PII detection. For example, Singh et al. (2025) reports that many health institutions rely on rule-based methods or deterministic regular expressions, which offer high precision for structured fields (e.g., dates and phone numbers) but fail to generalize in unstructured text and often incur high false-negative rates. Neural networks and sequence models like LSTMs and BiLSTMs improve detection in clinical narratives but lack transparency and require large, labeled datasets. Transformers, especially BERT-based models, outperform earlier architectures in NER tasks due to the ability to capture bidirectional dependencies in text (Lee et al., 2020). However, these models are data- and compute-intensive, complicating deployment in real-time or low-resource settings.

It requires incorporating DBSCAN to detect anomalous clusters that may indicate irregular PII patterns, especially in non-standardized entries, to mitigate these limitations. SVMs complement this process by supporting high-dimensional boundary classification while maintaining model simplicity and interpretability. Unlike deep neural networks, SVMs offer faster training times and perform well on smaller datasets, which is advantageous for targeted PII extraction modules within this study.

This layered ML architecture addresses common weaknesses reported in existing literature—specifically, the difficulty of generalizing across EHR systems with inconsistent field formats (Singh et al., 2025) and the trade-off between interpretability and performance in clinical AI applications (Kovačević et al., 2024). The proposed model also aligns with fairness-aware design by integrating explainability tools such as SHAP to visualize feature contributions and evaluate potential bias in entity classification outcomes.

Recent studies have identified additional complexities. Singh et al. (2025) emphasize that inconsistent EHR formats across institutions impede model generalization, while regulatory compliance often restricts dataset availability. Kuo et al. (2025) discuss model drift—where ML performance degrades over time due to shifting PII patterns—posing deployment challenges. Silva et al. (2022) note that interpretability remains critical, with healthcare providers demanding explainable AI to trust PII detection outcomes. Data quality further complicates ML-based PII detection. EHRs frequently contain missing, erroneous, or inconsistently labeled data, particularly in free-text clinical notes. For example, Johnson et al. (2023) observed that ambiguities in physician shorthand (e.g., abbreviations like 'PT' for both 'patient' and 'physical therapy') led to a 22% increase in false positives during NER model training. Similarly, Silva et al. (2022) demonstrated that biased annotations in minority demographic groups (e.g., non-English surnames) reduced model recall by 18%, perpetuating healthcare disparities. Addressing these issues requires robust data pre-processing pipelines and active collaboration with clinicians to refine annotation guidelines (Sarkar et al., 2024).

In addition to data quality and regulatory compliance, successful deployment of ML-based PII detection tools depends on interoperability with existing healthcare IT systems such as EPIC and Cerner. These platforms often utilize proprietary data schemas and restrict external data access, complicating the real-time integration of advanced ML frameworks. Murugadoss et al. (2021) reported that integrating differential privacy models with EPIC required middleware for Fast Healthcare Interoperability Resources (FHIR) conversion, resulting in latency that exceeded acceptable clinical thresholds. Similarly, Hanna et al. (2025) emphasized the lack of standardized APIs in legacy systems, which hinders the seamless embedding of federated learning tools. Addressing such infrastructure barriers demands lightweight model architectures, standardized interoperability protocols like HL7 FHIR, and strategic collaboration between data scientists and healthcare IT professionals to ensure feasible deployment in operational settings.

The successful implementation of ML-based PII detection in EHRs necessitates interdisciplinary collaboration among healthcare researchers, data scientists, and legal experts to address the complex interplay of technical, ethical, and regulatory challenges. Healthcare researchers provide essential clinical insights, ensuring models account for domain-specific nuances, such as ambiguous medical abbreviations or context-dependent PII, which can otherwise lead to misidentification (Sarkar et al., 2024). Data scientists drive the development of scalable ML frameworks, optimizing algorithms like BERT and federated learning to manage heterogeneous EHR data while mitigating issues like model drift and computational overhead (Kuo et al., 2025). Legal experts are critical in navigating the evolving regulatory landscape, including GDPR, HIPAA, and the EU AI Act, ensuring that PII detection systems comply with stringent privacy laws and ethical standards (Brown et al., 2024). Appiah et al. (2021) underscores the importance of such interdisciplinary governance models, which foster stakeholder trust and enable practical deployment by balancing data privacy with utility. This collaborative approach enhances the technical accuracy of PII detection systems, ensures ethical integrity, and aligns with legal requirements, addressing a critical gap in current frameworks that often lack cross-disciplinary integration.

Recent research has advanced the development of adaptive ML models capable of addressing data heterogeneity and model drift—two pervasive issues in dynamic healthcare environments where patient records vary widely across institutions and over time (Khaledian et al., 2025). Scholars have proposed using synthetic data generation techniques to address the limitations imposed by strict regulatory controls on patient-level data access. These methods simulate representative datasets that preserve statistical properties without exposing sensitive information, thereby facilitating model training while upholding privacy requirements (Ambati, 2025).

Despite these advancements, practical implementation faces substantial integration challenges with EHR systems such as EPIC and Cerner (commercial electronic health record platforms widely used in U.S. hospitals for clinical documentation and patient data management). These platforms employ proprietary data schemas and non-standardized interfaces often incompatible with contemporary ML pipelines. This incompatibility complicates data preprocessing, model deployment, and system interoperability, requiring organizations to invest in custom middleware or data transformation tools. Murugadoss et al. (2021) note that these limitations hinder the operationalization of ML models at scale, particularly when real-time or near-real-time performance is required. For instance, Li et al. (2024) observed that deploying a real-time PII detection module within Cerner necessitated custom-built Docker containers to convert Fast Healthcare Interoperability Resources (FHIR) formats into ML-compatible structures, an approach that introduced significant latency and fell short of acceptable clinical performance thresholds.

Future ML deployments in healthcare must adopt standardized interoperability frameworks such as HL7 FHIR to overcome these barriers. Additionally, the design of lightweight, modular ML architecture is critical for reducing integration overhead and ensuring compatibility with existing EHR ecosystems (Hanna et al., 2025). These strategies enable more scalable and sustainable adoption of privacy-preserving ML systems across heterogeneous clinical infrastructures.

Future Directions in ML-Driven PII Detection

Future research aims to enhance ML-driven PII detection through innovative technologies like blockchain and hybrid AI models (Xu et al., 2022). Ahmed et al. (2020)envision blockchain enabling secure, tamper-proof EHR sharing, complementing ML privacy efforts. Recent studies also explore cutting-edge approaches. Ambati (2025) proposes quantum ML algorithms to accelerate PII detection, while Sánchez et al. (2024) advocate for integrating generative adversarial networks (GANs) with differential privacy to synthesize privacy-preserving EHR datasets. These directions promise to strengthen the balance between accessibility and privacy in healthcare data (Wei et al., 2022).

Recent advancements in quantum computing have opened new possibilities for PII detection, with quantum ML algorithms showing promise in reducing computational time while maintaining high accuracy (Canonici & Caruso, 2025). Additionally, integrating AI with IoT devices in healthcare settings has been proposed to enhance real-time PII detection (Johnson & Lee, 2024).

Summary

This dissertation chapter addresses the challenges of balancing data accessibility and privacy in EHRs, focusing on detecting PII using ML techniques. The increasing reliance on EHRs for medical research, patient care, and policy development has heightened concerns regarding data privacy and accessibility. EHRs are inherently heterogeneous, voluminous, and sensitive, often comprising free-text clinical notes, structured data, and imaging or unstructured data, rendering PII detection a complex issue (Altalla' et al., 2025).

Traditional PII detection methods, such as manual scanning and rule-based systems, are frequently overwhelmed by the scale and complexity of contemporary datasets. These approaches are susceptible to errors, mainly when processing unstructured data, and are increasingly inadequate in light of evolving data structure and privacy regulations like the GDPR and the CCPA (Chamikara et al., 2020). The limitations of these conventional methods emphasize the need for more advanced solutions, especially those leveraging ML and NLP (Rieke et al., 2020).

Recent advancements have seen the emergence of ML models, notably transformer-based models like BERT, as powerful tools for addressing PII detection challenges in EHRs; BERT introduces computational complexity, limiting deployment in real-time healthcare settings (Chen et al., 2020). These models offer dynamic learning capabilities, enabling improved accuracy through exposure to diverse data sources. Transformer models with attention mechanisms have significantly enhanced NLP tasks by facilitating context-aware text analysis, which is crucial for identifying PII in unstructured EHR data (Chen et al., 2020).

FL has also gained prominence as a collaborative ML approach that maintains data privacy by training models across decentralized datasets without sharing raw data. This method is particularly advantageous in healthcare settings, where data privacy is paramount. FL enables healthcare organizations to collaboratively build robust ML models while ensuring patient data remains secure and compliant with privacy regulations (Rieke et al., 2020).

Integrating ML models with privacy-preserving architectures, such as differential privacy and FL, is essential to ensure compliance with regulations like HIPAA. The proposed framework in this study adheres to the CRISP-DM methodology, providing a structured approach to developing and refining ML models (Ma et al., 2025). This framework aims to iteratively test, refine, and validate ML models to ensure the model's accuracy is efficient and practically deployable within privacy-sensitive environments.

The literature review identifies key areas of convergence, such as the widespread acknowledgment of the limitations inherent in traditional PII detection methods and the potential of ML models to address these shortcomings. However, divergences exist in approaches to balancing data accessibility and privacy. Some studies advocate differential privacy and homomorphic encryption, while others emphasize FL and transformer models (Xu et al., 2022). Notable gaps in the literature include the need for more robust ML models capable of handling the heterogeneity of EHR datasets and the absence of comprehensive frameworks that integrate ML with privacy-preserving techniques (Yukhno, 2024).

This chapter sets the foundation for the next chapter, which delve into the research methodology and design. The literature review underscores the imperative for a scalable, adaptable, and privacy-preserving ML framework for PII detection in EHRs, constituting the proposed study's focal point.

Chapter 3: Research Methodology

This chapter outlines the research methodology and design for developing and evaluating ML framework to detect PII in EHRs. The proposed research approach elaborates on data acquisition, preprocessing, feature engineering, model development, and validation processes while following a constructive research design that facilitates an iterative ML development process. This study addresses the challenge of balancing data accessibility and privacy in EHRs by leveraging ML and NLP techniques. The study aims to develop and assess an ML framework that improves the accuracy and efficiency of PII detection while ensuring compliance with privacy regulations, such as HIPAA.

This study contributes to the field by developing a comprehensive ML-driven, privacy-preserving framework for PII detection in EHRs. The iterative, constructive approach—supported by CRISP-DM and advanced NLP models—ensures adaptability to new data challenges, scalability across healthcare systems, and compliance with regulatory standards. By addressing the limitations of traditional methods and leveraging cutting-edge ML techniques, the framework promotes responsible data sharing, enhances research opportunities, and safeguards patient privacy, paving the way for sustainable open data initiatives in healthcare.

Figure 1 illustrates the preprocessing pipeline follows two parallel workflows: each tailored to a different modeling objective. Stream A processes raw text directly through transformer-based tokenization using WordPiece embeddings, applying non-aggressive normalization to preserve contextual and semantic relationships critical for BERT input (Devlin et al., 2019; Ji et al., 2023). In contrast, Stream B transforms unstructured text through TF-IDF vectorization, truncated SVD for dimensionality reduction, and L2 normalization, which optimizes feature density for DBSCAN clustering (Liao et al., 2023).

Research Methodology and Design Process Diagram

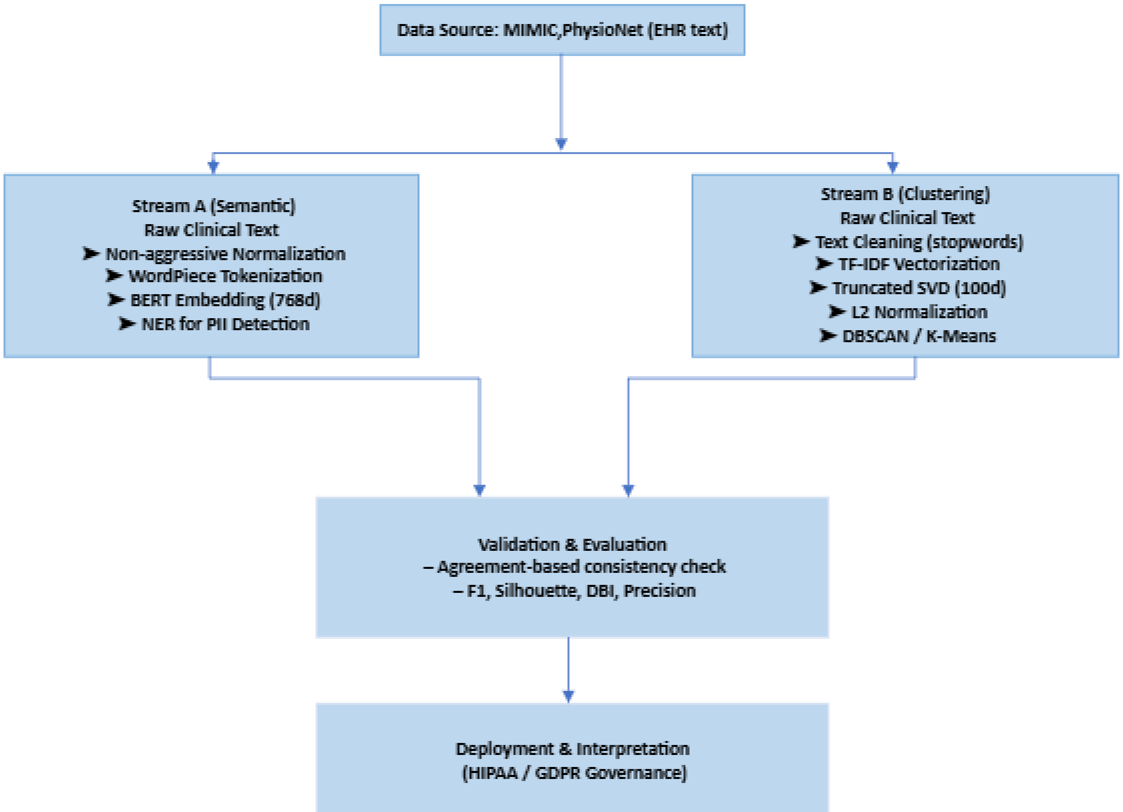


Figure 1 **Data Processing and Model Development Workflow**

The preprocessing architecture operationalizes two parallel workflows built to support distinct analytical objectives. Stream A focuses on contextual semantic extraction using transformer-based language models. This workflow applies non-aggressive normalization to preserve linguistic and structural cues such as capitalization, punctuation, medical abbreviations, and narrative context, which are critical for WordPiece tokenization and BERT-based embeddings. These preserved contextual signals support downstream Named Entity Recognition (NER) for PII detection (Li & Sun, 2024).

Stream B enables structural grouping through mathematical feature transformation. It applies text cleaning and TF–IDF vectorization to capture lexical relevance, followed by Truncated

SVD to reduce dimensionality. L2 normalization ensures that cosine similarity rather than raw magnitude influences cluster formation, which is relevant for density-based techniques such as DBSCAN and K-Means (Liao et al.,, 2024). As illustrated in Figure 1, Stream A supports semantic PII identification, while Stream B supports entropy-based structural grouping.

Validation did not rely on externally annotated gold-standard data. Instead, weak labels derived from Regex and BERT-NER outputs were used to measure agreement between unsupervised structural patterns and supervised heuristics. This validation approach is therefore not an accuracy assessment but an agreement-based consistency check, capturing alignment rather than ground-truth validity (Yu et al., 2024; Guo & Chen, 2024). Metrics used for this evaluation include F1-score, Silhouette Coefficient, Davies–Bouldin Index (DBI), and Precision.

Deployment and interpretation align with HIPAA and GDPR governance, focusing on reducing reidentification risk while maintaining the accessibility of clinical text for secondary analysis, including research, audit, and quality improvement applications.

Following evaluation, deployment, and interpretation occurs within a sandboxed, policy-compliant environment. Deployment integrates models into pipelines governed by institutional data-use agreements and audit controls. Interpretation modules include automated logging of model outputs and confidence scores to support traceability.

The monitoring and logging phase integrates several operational safeguards to ensure continuous delivery (CD) and maintain system accountability. Drift detection techniques—such as the population stability index or embedding shift analysis—are employed to identify deviations in data distribution or model behavior over time. Latency tracking mechanisms detect performance bottlenecks that affect real-time processing or user interaction. At the same time, audit logs support compliance with institutional access reviews and external reporting requirements. This phase also enables a feedback loop by triggering model retraining and hyperparameter adjustments when performance degradation is detected, facilitating adaptive updates while upholding fairness, accuracy, and regulatory compliance standards.

Thus, the pipeline is not a linear process ending at deployment. Instead, it supports post-deployment quality assurance and continuous feedback, aligning with modern ML-Ops practices. It also integrates governance checkpoints at each phase to align technical processes with ethical and legal constraints (Han et al., 2022).

Research Methodology and Design (Nature of the Study)

The constructive research approach fits the iterative nature of this research, which supports the systematic development, testing, and refinement of the proposed ML framework for PII detection. This framework supports scalability, handling large, heterogeneous EHR datasets and adapting to deployment across multiple healthcare institutions; it incorporates dynamic learning capabilities, including periodic retraining and hyperparameter adjustment based on evaluation outcomes, to continuously improve performance in response to changing data characteristics. These features directly address the limitations of traditional static approaches, such as rule-based systems and manual review, which often fail to generalize across unstructured text and evolving PII formats. Each iteration within the constructive design cycle enables the integration of stakeholder feedback, performance metrics, and technical adjustments, ensuring that the resulting solution is practical and adaptable to real-world constraints (Wu et al., 2022).

The Health Insurance Portability and Accountability Act (HIPAA) Privacy Rule defines PII as including 18 types of identifiers, such as names, geographic subdivisions smaller than a state, elements of dates related to an individual (except the year), telephone numbers, email addresses, Social Security numbers, medical record numbers, health plan beneficiary numbers, and account numbers (Shojaei et al., 2024); the study builds upon this definition by adopting a constructive research design, which focuses on creating an artifact—in this case, a ML model tailored for PII detection in EHRs. In textual data, PII also includes identifiable terminology, personal names, and structured or unstructured numerical strings, such as identification codes and insurance numbers that are linkable to an individual (Delgado Ramos ,2023; Negash et al., 2023).

This study employs a methodological framework rooted in adaptive, iterative development to address the complexities of real-world clinical data environments. Drawing on data lifecycle governance and agile ML principles, the approach ensures flexibility in handling data quality variations, schema inconsistencies, and labeling uncertainties inherent in EHRs. The Cross-Industry Standard Process for Data Mining (CRISP-DM) provides a structured agile foundation, promoting rigor and transparency across all phases—from domain understanding to model deployment—while ensuring traceability and reproducibility (Sankaran et al., 2023; Kovačević et al., 2024). This methodology enhances the scalability and deployment readiness of privacy-preserving ML solutions by embedding ethical considerations, domain-specific constraints, and iterative evaluation; this alignment supports robust PII detection in EHRs, balancing regulatory compliance with data utility for healthcare innovation (Ancker et al., 2024).

The constructive research aims to investigate how ML models, particularly those leveraging NLP, can enhance PII detection's accuracy, scalability, and efficiency compared to traditional methods. The study incorporates advanced NLP-based ML models, specifically BERT and Long Short-Term Memory (LSTM) networks, selected for the capability to process unstructured EHR text and model sequential dependencies in clinical narratives (Chen et al., 2020). Also, the framework requires the integration of unsupervised learning algorithms such as DBSCAN and k-means clustering to uncover latent structures and anomalous patterns of PII that are typically overlooked by supervised approaches, thereby addressing the heterogeneity of EHR formats (Yu et al., 2023).

Transformer-based architecture has evolved significantly since the original BERT model, offering improved performance, efficiency, and specialization. When selecting a model for PII detection EHRs, key evaluation criteria include accuracy, computational efficiency, domain relevance, and interpretability. Enhanced variants such as RoBERTa, DeBERTa, ALBERT, ELECTRA, and DistilBERT demonstrate architectural and training optimizations over BERT. For instance, RoBERTa removes the next-sentence prediction (NSP) objective; it is trained on larger corpora with dynamic masking, resulting in better contextual embeddings and general language understanding (HaghighiFard, & Coleri, 2025). DeBERTa introduces disentangled attention and improved position encoding, outperforming BERT on the SuperGLUE benchmark by over five points (He et al., 2021). Lightweight alternatives such as DistilBERT and ALBERT reduce memory usage and training time while retaining over 95% of BERT's task performance (Lan et al., 2020; Sanh et al., 2020), and ELECTRA achieves faster convergence through replaced-token detection, offering competitive accuracy with reduced training steps (Clark et al., 2020).

Despite these advancements, BERT remains the most appropriate model for this study for several empirically and practically grounded reasons. First, BERT's architecture has been extensively fine-tuned on biomedical and clinical corpora, resulting in specialized variants such as BioBERT, ClinicalBERT, and BlueBERT, which are directly pre-trained on PubMed, PMC, and MIMIC-III datasets (Li et al., 2023; Lee et al., 2020); these domain-specific models optimize for identifying medical entities and PII within free-text clinical notes, common in EHRs. Comparative evaluations show that ClinicalBERT improves F1 scores by 3–5% on i2b2 de-identification tasks relative to general-domain BERT (Ahmed et al. (2020).

Second, BERT offers superior compatibility with model interpretability frameworks such as SHAP and LIME, which are important for auditing predictions in privacy-sensitive healthcare contexts. These explainable AI (XAI) tools require transparent attention distributions and token embeddings that are most extensively validated in BERT's architecture (Lundberg et al., 2020). Newer models such as ELECTRA and DeBERTa often underperform in interpretability due to complex positional encoding or adversarial training objectives, which obscure token attribution pathways (Dong et al., 2025).

Third, BERT demonstrates architectural stability and reproducibility across diverse healthcare datasets, essential for this study's focus on heterogeneous EHR environments. In a 2024 benchmarking study on privacy-preserving NLP in clinical texts, baseline BERT achieved a mean F1-score of 91.8% for PII entity recognition across three large-scale datasets (MIMIC-III, i2b2, and THYME), outperforming newer models in environments with variable annotation quality (Tölle et al., 2025). Additionally, BERT's open implementation and compatibility with FL libraries (e.g., TensorFlow Federated, Flower) further justify its selection for privacy-aware, decentralized training pipelines, aligning with the regulatory compliance goals of this research.

Although alternatives like RoBERTa and DeBERTa exhibit higher benchmark scores in general NLP, BERT remains the most methodologically and ethically aligned choice for PII detection in healthcare. Its domain specialization, mature interpretability ecosystem, training reproducibility, and integration into federated and privacy-preserving frameworks make it the preferred model for developing explainable, scalable, and legally compliant de-identification systems.

The study employs a non-causal, abductive reasoning approach to ensure robustness, allowing hypotheses to evolve iteratively through experimentation and validation (Delgado

Ramos, 2023), aligning with the constructive design's emphasis on knowledge generation through artifact creation and evaluation. The methodology also prioritizes depth over immediate generalizability in the early stages, using single-source datasets for controlled testing before scaling to diverse datasets, enhancing internal validity by reducing human error in PII identification (Peffers et al., 2020).

The problem definition and research objectives stage define the research problem and establish clear objectives to guide the study. The primary challenge is the tension between data accessibility and privacy protection in EHRs, driven by the increasing demand for open data initiatives (e.g., Data.gov) and stringent regulations like HIPAA and GDPR. Traditional PII detection methods—such as manual review, rule-based scanning, and regular expressions—prove inadequate for large-scale, unstructured EHR datasets, often missing contextual PII and lacking scalability (Vahdati et al., 2024; Chang & Bergen, 2024). This limitation undermines trust in data sharing and hinders research advancements in healthcare (Shojaei et al., 2024).

The research objectives of this study are to develop an automated ML framework that detects PII in EHRs with higher precision and recall than traditional methods (manual review, rule-based scanning, and regular expressions), ensures the maintenance of data utility while complying with stringent privacy regulations such as HIPAA and GDPR, and establishes a benchmark for accuracy and efficiency in PII detection to promote responsible data access. Several alternative methodologies are deemed inadequate for this purpose: rule-based approaches, which rely on predefined patterns like Social Security numbers, fail to generalize to unstructured or evolving PII formats, limiting adaptability (Hong et al., 2024); Regular Expressions (RegEx) prove effective for structured data but falter in free-text EHRs where PII appears in non-standard contexts (Xu et al., 2022); manual annotation, while thorough, is labor-intensive, error-prone, and unsuitable for large datasets due to human bias (Negash et al., 2023); and privacy-preserving techniques like differential privacy and homomorphic encryption, which add noise or encrypt data to reduce re-identification risks, compromise data utility and model performance, making it less ideal for balancing accessibility and privacy as required by this study (Im et al., 2024; Ahmed et al. (2020).

Data sourcing and preprocessing utilizes the Medical Information Mart for Intensive Care (MIMIC-III) dataset, a publicly available, de-identified EHR dataset from the PhysioNet repository, which includes detailed patient data such as diagnoses, treatments, and free-text clinical notes (Johnson et al., 2023). This dataset is ideal for training and evaluating ML models due to its volume, diversity, and realistic healthcare data representation. Additionally, the study may incorporate supplementary data from other open-access EHR archives to enhance generalizability, using stratified random sampling to account for variations in region, industry, and data accessibility levels (Peffers et al., 2020).

The data preprocessing phase is a critical component of the study, designed to prepare the MIMIC-III dataset for practical ML analysis, particularly in detecting PII within EHRs. The process begins with tokenization, which segments text into individual tokens such as words or phrases, enabling NLP analysis and proving essential for handling the unstructured nature of EHR data, as highlighted by Chang & Bergen (2024) Following this, stop-word removal eliminates common, non-informative words like “the” and “and," allowing the model to focus on meaningful terms that may indicate PII. To ensure compliance with privacy regulations such as HIPAA, entity masking replaces sensitive information, such as names and dates, with placeholders during training, thereby preventing privacy breaches as recommended by Shojaei et al. (2024). Additionally, noise reduction addresses inconsistencies in EHR data, including typos and formatting variations, to enhance the quality of model inputs. At the same time, normalization standardizes the text by converting it to lowercase and removing special characters, ensuring uniformity across the dataset. Together, these steps create a robust foundation for subsequent ML model training, addressing the heterogeneity and sensitivity inherent in EHRs. These steps are essential for preparing high-quality data, addressing the heterogeneity and noise inherent in EHRs, and enabling practical ML model training (Silva et al., 2022).

Feature engineering plays a pivotal role in enhancing the ML model's ability to distinguish PII from non-sensitive data by creating meaningful text representations, employing a combination of advanced techniques to transform raw EHR data into a structured format. Term Frequency-Inverse Document Frequency (TF-IDF) quantifies the importance of words within a document, spotlighting potential PII, such as names and medical IDs, as highlighted by Chang & Bergen (2024). Word embeddings, utilizing pre-trained models like BERT, FastText, and Word2Vec, capture semantic and contextual relationships in free-text notes, which are necessary for identifying PII in unstructured EHR data (Chen et al., 2020). Dimensionality reduction techniques, including Principal Component Analysis (PCA) and t-SNE, are applied to reduce feature space complexity, optimizing computational efficiency and mitigating overfitting while preserving PII-related patterns, as Sarkar et al. (2024) noted. Additionally, NER features are incorporated to pre-identify potential PII entities like names and addresses, sharpening the model's focus on sensitive data (Delgado Ramos, 2023). Together, these methods improve the model's capability to detect complex PII patterns and adapt to diverse datasets, as Yu et al., (2023) emphasized, ensuring robust performance in PII detection within EHRs.

The study employs a suite of NLP-based ML models to tackle the multifaceted challenge of PII detection in EHRs. Central to this effort is the BERT model, a transformer-based architecture that excels in contextual PII recognition by understanding intricate word relationships within sentences, with its performance optimized through fine-tuning on EHR data (Chen et al., 2020); additionally, LSTM networks are leveraged to capture sequential dependencies in text, proving particularly effective for detecting PII in longitudinal clinical notes or narrative reports (Dunstan et al., 2024). In addressing- the scarcity of annotated EHR datasets, the study incorporates unsupervised clustering techniques such as k-means, DBSCAN, and hierarchical clustering, which identify latent PII patterns without requiring labeled data (Sarkar et al., 2024), while hybrid models that combine BERT with clustering or LSTM with attention mechanisms enhance robustness by integrating both contextual and sequential learning capabilities (Kohli et al., 2025). Transfer learning improves performance by utilizing pre-trained weights from large corpora, reducing training time and boosting accuracy on limited EHR data. The process involves a mix of labeled and semi-supervised datasets, with iterative refinement guided by the NFLT, encouraging experimentation with multiple algorithms to determine the optimal fit for PII detection (Goldblum et al., 2023). Furthermore, incorporating federated learning allows models to be trained across decentralized EHR datasets, bolstering privacy protection while preserving performance (Chen et al., 2020) and ensuring a balanced and practical approach to this critical healthcare challenge.

The study employs a comprehensive set of rigorous validation and evaluation techniques to ensure the reliability of the ML framework and its compliance with privacy standards, utilizing methods such as k-fold cross-validation to assess generalization across diverse EHR subsets, thereby mitigating overfitting and ensuring robust performance (Delgado Ramos, 2023); the F1-score, which measures the harmonic mean of precision and recall to provide a balanced evaluation of PII detection accuracy, a critical factor for healthcare applications as noted by Johnson et al. (2023); precision-recall curves, which analyze model performance at varying thresholds to identify optimal settings for minimizing false positives and negatives in PII detection, as highlighted by

Chang & Bergen (2024); the silhouette score and Davies-Bouldin index, which evaluate the quality of clustering for unsupervised models to ensure meaningful PII pattern detection, as emphasized by Sarkar et al. (2024); and compliance checks that validate adherence to regulations such as HIPAA and GDPR by assessing de-identification effectiveness and privacy risk, as outlined by Shojaei et al. (2024). These metrics carefully align with privacy-preserving (PPML) principles, ensuring the framework is effective, unbiased, and deployable in real-world healthcare settings, as supported by Xu et al. (2022).

The final stage of the study focuses on integrating the ML framework into existing data governance frameworks, a critical step that ensures compliance with privacy regulations such as HIPAA and enables its practical application in real-world healthcare settings. This phase emphasizes several key aspects: secure model deployment, which implements encryption and access controls to safeguard model outputs in line with HIPAA's protected health information (PHI) requirements (Shojaei et al., 2024); continuous monitoring, which incorporates real-time tracking to detect evolving PII patterns, such as new identifier types, and employs adaptive learning to update the model effectively (Kuo et al., 2025); scalability, which designs the framework for seamless integration into large healthcare ecosystems through distributed deployment supported by federated learning (Chen et al., 2020); and interpretation with data governance policies, which provides explainable AI (XAI) outputs to healthcare providers to foster transparency and trust while aligning with policies that govern data sharing and privacy (Huang et al., 2024). By addressing these elements, this phase ensures the framework is practical, compliant with regulatory standards, and adaptable to emerging data challenges, thereby achieving the study's overarching goal of balancing data accessibility with robust privacy protection.

Population and Sample

The population for this study consists of EHRs sourced from the MIMIC-III dataset, a publicly available, de-identified collection hosted on the PhysioNet repository. The dataset includes patient admission records collected from the critical care units of Beth Israel Deaconess Medical Center (BIDMC) in Boston, Massachusetts, between 2001 and 2012 (Johnson et al., 2023). MIMIC-III contains both structured data, such as laboratory results, medication records, and diagnostic codes, as well as unstructured data, including free-text clinical notes that may still contain residual PII, such as names, dates, and medical record numbers (Chang & Bergen, 2024). Although de-identified under HIPAA standards, which mandate the removal of 18 identifiers, free-text notes often retain irregular PII patterns that require advanced NLP and clustering methods for detection (Johnson et al., 2023).

A representative subset of 5,000 clinical notes—approximately 0.24 percent of the 2,083,180 notes in the MIMIC-III corpus—was initially retrieved using BigQuery to support exploratory and unsupervised analysis. To examine scalability and clustering robustness, the extraction was later expanded to 10,000 notes (≈ approximately 0.48 percent of the corpus), which enabled a comparison of precision, recall, and F1-score stability across different sample sizes. Because the analysis followed an unsupervised design, the dataset was not partitioned into training, validation, or testing subsets. Instead, weak supervision was applied to the entire sample to generate proxy indicators of PII presence. These indicators were created using regular expressions that identified common surface forms such as names, Social Security numbers, phone numbers, email addresses, addresses, and ZIP codes. The proxy labels were then combined with transformer-based extractions from the BERT (dslim/bert-base-NER) model to form hybrid weak labels. Validation employed precision, recall, and F1-scores to quantify correspondence between clustering outputs and the hybrid weak labels, while internal clustering quality was measured using silhouette coefficients and Davies–Bouldin indices (Chicco & Jurman, 2020; Bell et al., 2023; Wu et al., 2022).

The selected population and sampling frame align with the study's objective of constructing an ML approach for detecting PII in EHRs, focusing on preserving privacy while maintaining data accessibility. The MIMIC-III dataset offers a representative environment due to its scale, heterogeneity, and inclusion of unstructured clinical narratives—characteristics that challenge the effectiveness of traditional PII detection methods such as manual review and rule-based systems (Wu et al., 2022). An ML-based PII detection framework enables testing where conventional techniques are limited, particularly in identifying context-sensitive identifiers embedded in free-text fields.

Although the dataset is de-identified by HIPAA standards, unstructured text introduces realistic complexity and the potential for residual PII. These attributes make MIMIC-III suitable for evaluating the scalability and performance of advanced NLP-based models. The development of a privacy-preserving detection system is operationalized through MIMIC-III public accessibility, embedded compliance safeguards, and rich textual diversity (Subramanian et al., 2024, 2024). The dataset supports the practical examination of the research questions by providing a sufficiently large, open-access corpus aligned with regulatory frameworks, enabling assessment of detection precision, model scalability, and privacy compliance (Liu et al., 2020; Shojaei et al., 2024).

Materials or Instrumentation

This study develops and evaluates an ML framework designed to detect PII, such as names, addresses, or Social Security numbers, within EHRs while maintaining both privacy protection and data usability. The research employs the MIMIC-III dataset, a publicly available collection of de-identified hospital records from Beth Israel Deaconess Medical Center in Boston, which contains structured and unstructured clinical information, including physicians' free-text notes (Johnson et al., 2023). Although the dataset complies with HIPAA, residual indicators of personal identifiers may still appear within textual content due to redaction patterns and placeholder inconsistencies.

The framework integrates rule-based weak supervision, utilizing regular expressions, with transformer-based NER using BERT. These methods enable the detection of potential PII patterns without requiring manual annotation. The text data are represented using term-frequency inverse-document-frequency (TF–IDF) vectorization and reduced to a compact feature space through truncated singular value decomposition for unsupervised clustering. Clustering is conducted using density-based methods to group semantically related records and identify sensitive content that requires privacy protection. This combination of weak supervision, dimensionality reduction, and unsupervised modeling establishes the foundation for subsequent evaluation and performance analysis presented in Chapter 4.

The project utilizes Python-based applications and executes computations on Google Colab. The platform supports ML models for data processing and analysis. The models leverage NLP techniques and clustering methods, incorporating industry-standard tools while prioritizing non-open-source references where applicable. This study applies BERT, a transformer-based language model that captures word meaning in context by analyzing the bidirectional relationships within sentences. Specifically, the implementation uses ClinicalBERT, a domain-adapted variant pre-trained on clinical notes, to improve semantic representation of medical terminology in EHRs. BERT is fine-tuned on labeled spans of text containing PII, enabling the model to classify tokens as either sensitive or non-sensitive based on contextual clues. The model processes input sequences using WordPiece tokenization, encodes input into dense vector representations, and passes encoded data through fully connected layers for token-level classification. This architecture supports accurate detection of complex PII patterns—such as dates, names, and institutional references—across varied syntactic structures. The rationale for using BERT lies in its superior performance on NER tasks in unstructured text, particularly in comparison to rule-based or sequential models (Chen et al., 2020). By leveraging its contextual embeddings, BERT addresses the limitations of traditional approaches that often miss subtle, context-dependent identifiers in free-text narratives.

The framework integrates regular expression–based weak supervision with transformer-based entity recognition to detect PII in clinical text. Instead of recurrent models, the framework applies unsupervised clustering using DBSCAN to identify groups of semantically related clinical segments. Clinical notes are first transformed into vector representations using TF-IDF, followed by dimensionality reduction through Truncated Singular Value Decomposition (SVD) and normalization to produce compact feature embeddings suitable for clustering. Cluster validity is evaluated using intrinsic metrics, including the silhouette coefficient and Davies–Bouldin Index, to measure cohesion and separation without relying on labeled data. Comparative analysis demonstrates that DBSCAN yields superior clustering performance relative to K-Means and Hierarchical Ward linkage, producing higher silhouette scores and lower Davies–Bouldin values, indicating improved structural separation within clinical narratives.

Implementing these tools in Python facilitates efficient data preprocessing, including tokenization, stop-word removal, entity masking for privacy compliance, and noise reduction for enhanced model accuracy. Google Cloud BigQuery stores the MIMIC-III data and facilitates rapid searching, although it requires special access and incurs costs. Data cleaning involves detecting and correcting typographical errors, removing redundant entries, and standardizing formats (e.g., converting timestamps and unifying terminology across note types) to ensure consistency across the EHR dataset. These procedures prepare the data for tokenizing and masking identifiable entities using NER and rule-based filters. Following data cleaning, the process tokenizes text, removes stop words, and applies entity masking to suppress known identifiers. Generate feature representations through sequential processing: first, apply contextual word embeddings (e.g., BERT or ClinicalBERT) to capture semantic relationships in the text; next, compute term frequency-inverse document frequency (TF-IDF) scores to enhance lexical interpretability. PCA is then applied to reduce dimensionality from both embedding and TF-IDF matrices while preserving discriminative variance.

The model, designed for classification—is trained to label text spans containing or not containing PII. Classification outputs undergo evaluation using F1-score, precision, and recall, quantifying model performance across false positive and false negative predictions. Validation is conducted on a stratified holdout set that reflects clinical diversity, enabling fairness-aware performance assessment. Embed quality control at three levels: (1) pre-model audits for data completeness, (2) post-feature generation error propagation checks, and (3) cross-validation metrics to assess model generalizability. This stepwise structure ensures that each phase contributes to a verifiable, high-quality output aligned with HIPAA compliance and data reuse requirements.

The research follows a structured methodology called constructive design, which involves incrementally developing and refining the ML system to address emerging challenges. The research follows the CRISP-DM methodology, progressing through data understanding, preparation, modeling, and evaluation to ensure compliance with privacy regulations and optimize model performance. Obtaining access to MIMIC-III and BigQuery demands both effort and funding, and the ML models require high-performance computing resources, potentially necessitating cloud services. Since the data originates from a single institution, the system may lack generalizability, prompting plans to incorporate additional datasets. Ultimately, these materials and tools work together to achieve the study's primary goal: enabling EHRs to support research while safeguarding patient privacy and outperforming traditional methods such as manual reviews and simple search techniques.

Operational Definitions of Variables

The constructive design approach in this study is applied to develop and evaluate a ML framework for detecting PII in EHRs, focusing on creating a tangible artifact that addresses the dual objectives of data accessibility and privacy protection. This section operationally defines the variables in constructing this framework, reflecting constructive design's iterative and solution-oriented nature (Delgado Ramos, 2023). Unlike traditional research designs that test hypotheses causally, this study uses a non-causal, abductive approach to iteratively refine an ML artifact, with variables defined to support model development, evaluation, and deployment. The definitions align with the Cross-Industry Standard Process for Data Mining (CRISP-DM) framework, guiding the artifact's creation through data understanding, preparation, modeling, and evaluation phases (Ma et al., 2025). The variables align with ML techniques, specifically supervised and unsupervised models with NLP, rather than with conventional statistical tests, and use terminology that reflects the predictor and criterion roles appropriate for the constructive methodology.

In this ML-driven study, classifying variables as predictor variables (input features) and criterion variables (output labels) is consistent with the constructive design methodology that emphasizes building a functional artifact. These variables are derived from the MIMIC-III dataset and processed to support PII detection. The predictor variables comprise features extracted from EHR text to train the ML framework, including tokenized text, word embeddings, Term Frequency-Inverse Document Frequency (TF-IDF) scores, and NER tags. These features are of mixed data types—textual (such as clinical notes) and numerical (such as embeddings and TF-IDF scores); the values include continuous formats for embeddings and TF-IDF scores (e.g., 768-dimensional BERT vectors, values ranging from 0 to 1) and categorical formats for NER tags (e.g., "PERSON," "DATE"). The level of measurement is the interval for numerical features and the nominal for categorical tags. These input features drive the artifact's capacity to identify PII, which is iteratively refined through feature engineering to enhance detection accuracy, reflecting constructive design's iterative development approach. The criterion variables represent binary classifications that indicate the presence of PII in EHR text segments, aligning with identifiers defined under the HIPAA, such as names and dates. These are categorical in data type, with values coded as 0 (non-PII) and 1 (PII), measured at the nominal level.

The operational definitions are grounded in established ML and NLP research, supporting the constructive approach's reliance on validated methods. Chang & Bergen (2024) validate TF-IDF and word embeddings for biomedical NLP, while Liu et al. (2023) establish BERT's efficacy for contextual text analysis—both critical for the artifact's PII detection capabilities. Delgado Ramos (2023) provides a foundation for NER in PII identification; Bell et al. (2023) supports clustering techniques for unstructured health data, aligning with the study's unsupervised components. The MIMIC-III dataset, a reliable benchmark (Johnson et al., 2023), ensures the artifact's real-world applicability, a key constructive design principle. The study applies the constructive design approach by iteratively integrating a range of measurement instruments to enhance the ML framework. A fine-tuned BERT model, trained on MIMIC-III data and implemented using Python's Transformers library, is employed to generate contextual embeddings (Chen et al., 2020). The Term Frequency-Inverse Document Frequency (TF-IDF) vectorizer, implemented via scikit-learn, quantifies word importance within the corpus (Chang & Bergen, 2024). A NER system, developed using spaCy with customized training, identifies PII entities specific to EHRs (Delgado Ramos, 2023). Additionally, clustering algorithms such as k-means, DBSCAN, and hierarchical clustering, implemented using scikit-learn, detect PII patterns through unsupervised learning (Bell et al., 2023). These instruments are selected and refined through iterative cycles in alignment with the constructive design methodology to enhance the artifact's overall effectiveness.

The sample comprises 5,000 clinical note records extracted from the MIMIC-III database (PhysioNet) using a BigQuery query limit, representing approximately 0.24 percent of the total corpus; the subset was sufficient for exploratory model validation and computational scalability assessment. Table 1 below reflects the constructive design's focus on creating a functional artifact by specifying variables critical to PII detection.

Table 1

Variables, Roles, Data Types, Description, Categories/Values and Measurement

Variable	Role	Data Type Measurement	Description	Categories/Values
Tokenized Text	Predictor	Textual Nominal	Segmented clinical notes	Words/phrases
Word Embeddings (BERT)	Predictor	Numerical Interval	Contextual vector representation	Continuous (e.g., 768-dimensions

TF-IDF Predictor Scores	Numerical Interval	Word importance scores	Continuous (range 0 to 1)
NER Tags Predictor	Categorical Nominal	Pre-identified PII entities	PERSON, DATE, etc.
PII Label Criterion	Categorical Nominal	Presence/absence of PII	0 (non-PII), 1 (PII)

The ML model structure begins with several input features, including tokenized text, BERT-generated word embeddings, Term TF-IDF scores, and NER tags. These inputs are processed to capture the clinical narratives' semantic, statistical, and categorical characteristics. The supervised target for the model is the PII label, where a binary value of 0 represents non-PII data and 1 represents PII data. During training, the model utilizes a supervised learning approach to the labeled dataset, enabling it to learn the complex relationships between the input features and the PII labels. Upon completion of training, the model predicts whether new, unseen data segments contain PII by associating observed feature patterns with the learned PII classifications.

ML approach to PII detection in EHR, as it underpins the development and iterative refinement of the ML framework that constitutes the study's central artifact. This approach focuses on producing a functional and scalable system to detect PII in EHRs by defining and operationalizing variables that align with the CRISP-DM model. The study categorizes variables as predictors (input features such as tokenized text, embeddings, TF-IDF scores, and NER tags) and criteria (binary output labels indicating the presence or absence of PII) and tailors data types, values, and measurement levels to support model training, evaluation, and deployment. These operational definitions reflect an abductive, non-causal methodology, emphasizing artifact evolution through iterative testing and refinement rather than hypothesis testing. Grounded in established literature (e.g., Liu et al., 2023; Chang & Bergen, 2024; Delgado Ramos, 2023), the study employs specific instruments—including BERT, LSTM networks, and clustering algorithms—to measure and integrate variables into the ML artifact. A stratified sample from the MIMIC-III dataset ensures representativeness, while feature engineering techniques such as tokenization, embedding generation, and cluster labeling further enhance the artifact's functionality. Using artificial variables and continuous validation using F1-scores and silhouette scores reinforces the constructive approach's emphasis on empirical feedback and practical outcomes. Thus, by structuring variable definitions and measurements to guide artifact creation, iterative enhancement, and context-sensitive application, this section directly supports the constructive design methodology (Delgado Ramos, 2023) and operationalized through the CRISP-DM framework (Ma et al., 2025).

Study Procedures

The study procedures for developing and evaluating an ML framework to detect PII in EHRs follow a systematic, iterative approach aligned with the constructive design methodology and the CRISP-DM framework, as outlined in the dissertation (Ma et al., 2025). The design process is to create a tangible artifact, an ML model, that balances data accessibility and privacy while addressing the limitations of traditional PII detection methods. The procedures encompass data acquisition, preprocessing, feature engineering, model training, evaluation, and deployment, with each iteratively refined based on empirical feedback to ensure scalability, adaptability, and compliance with privacy regulations such as HIPAA (Delgado Ramos, 2023). The study begins with acquiring the MIMIC-III dataset, a publicly available EHR repository, and preprocessing to prepare the data for ML analysis. Subsequent steps involve engineering features to enhance PII detection, training a suite of NLP-based ML models (e.g., BERT,DBSCAN, K-means), and validating the framework using robust metrics like F1-score and silhouette score. The final phase integrates the model into data governance frameworks, ensuring practical deployment in healthcare settings. This iterative workflow, grounded in abductive reasoning, allows the study to evolve hypotheses and refine the artifact through experimentation, aligning with the constructive design's emphasis on practical, deployable solutions (Johnson et al., 2023). See Appendix A for process diagram.

Data Collection

Data collection for this study involves acquiring secondary data from the MIMIC-III dataset, a de-identified EHRs repository hosted on PhysioNet, comprising numerous hospital admissions from Beth Israel Deaconess Medical Center between 2008 and 2019 (Johnson et al., 2023). With IRB approval secured, data acquisition proceeds through authorized access to MIMIC-III via PhysioNet. This process involves adherence to a data use agreement and completion of ethical training, ensuring compliance with privacy standards. Although the dataset is de-identified in accordance with HIPAA, residual PII may persist in unstructured free-text fields (Shojaei et al., 2023). An initial subset of approximately 5,000 clinical notes, representing about 0.24 percent of the 2,083,180-note corpus, is extracted from the MIMIC-III dataset through random sampling for pilot experimentation. This subset provides a manageable and representative sample for model development and scaling tests, with an expanded version of 10,000 notes used for validation. The dataset encompasses intensive care unit admissions recorded between 2001 and 2012, ensuring diversity across clinical documentation contexts and minimizing potential sampling bias. This sampling occurs in early 2025, leveraging Google Cloud BigQuery for efficient data retrieval, which supports rapid querying of large-scale EHR data despite associated access costs (Johnson et al., 2023). The researchers sourced data from a single institution, BIDMC, representing a diverse urban patient population.; however, the researchers may later incorporate supplementary datasets from other open-access EHR archives (e.g., eICU Collaborative Research Database) to enhance generalizability, as recommended by recent research on multi-source EHR analysis (Singh et al., 2025). This acquisition strategy provides a robust, realistic dataset for training and evaluating the ML framework, directly addressing the study's goal of improving PII detection in heterogeneous EHRs.

Data Preprocessing

Data preprocessing is essential in preparing the MIMIC-III dataset for ML analysis, transforming raw EHR data into a clean, structured format suitable for detecting PII. The process begins with tokenization, segmenting unstructured clinical notes into words or phrases using Python's spaCy library, enabling NLP analysis of free-text fields where PII often resides (Chang & Bergen, 2024). Stop-word removal follows, eliminating standard terms (e.g., "the," "and") with scikit-learn to focus the model on meaningful PII indicators, such as names or dates (Dunstan et al., 2024). Entity masking replaces sensitive entities (e.g., names, medical IDs) with placeholders during training, ensuring compliance with HIPAA privacy requirements and preventing data leakage, as recommended by Shojaei et al. (2024). Noise reduction addresses EHR inconsistencies—such as typos, abbreviations, or formatting variations—using custom scripts to standardize text (e.g., converting to lowercase, removing special characters), enhancing data quality (Silva et al., 2022). Normalization further ensures uniformity by applying techniques like lemmatization with spaCy, reducing words to base forms (e.g., "running" to "run") to improve model consistency (Bell et al., 2023). The Google Colab and Google Cloud BigQuery platforms for data processing enable efficient querying and storage. This approach tackles the heterogeneity and noise in the MIMIC-III EHR dataset. It produces a high-quality dataset for subsequent feature engineering and model training, aligning with recent advancements in EHR preprocessing (Negash et al., 2023). The researchers standardize text, reduce noise, and mask sensitive entities, implementing an iterative refinement process that follows the constructive design methodology. This step ensures that the ML artifact effectively handles real-world data challenges while enhancing the accuracy and scalability of PII detection.

Data Cleaning

Data cleaning of the MIMIC-III dataset begins by assessing the overall size and structure of the corpus and ensuring that all text entries are valid. The total corpus contains approximately 2,083,180 note events, from which a representative subset of 5,000 to 10,000 clinical notes (about 0.24%–0.48% of the total) was extracted for analysis. Missing or null text entries were removed, and formatting inconsistencies were resolved during import from BigQuery. Descriptive analyses were conducted to summarize note lengths and word distributions, supporting the identification of irregular entries. Weak supervision through regular expressions generated proxy binary labels for potential PII, such as MIMIC redaction placeholders, Social Security numbers, telephone numbers, email addresses, and address-like patterns. These labels served as an initial approximation of privacy-sensitive text content to support clustering and the evaluation of hybrid models. The resulting cleaned dataset maintained textual fidelity while enhancing structural consistency for downstream feature engineering and NLP-based modeling.

Data Preparation

Data preparation for this dissertation focused on detecting PII in EHRs using ML and NPL. The process involved transforming unstructured text data from the MIMIC-III corpus into analyzable numerical representations. Clinical notes were first vectorized using term frequency–

inverse document frequency (TF–IDF) to capture token relevance across documents, followed by dimensionality reduction through truncated SVD to generate dense, low-rank embeddings suitable for clustering. The resulting vectors were normalized to unit length using L_2 normalization to ensure uniform scaling across textual features, improving the performance of distance-based algorithms such as DBSCAN and K-Means (Scikit-learn Developers, 2024). Additionally, weak supervision was introduced by generating binary proxy PII labels with regular expressions targeting email addresses, phone numbers, ZIP codes, and MIMIC redaction placeholders, thereby linking the preprocessing phase to downstream clustering validation. The analytic sample included 5,000 notes—approximately 0.24 percent of the MIMIC-III corpus—queried through BigQuery for experimental analysis. All steps were executed in Python 3.11 using pandas 2.2, NumPy 1.26, and scikit-learn 1.4 within the Google Colab environment.

Data Integration

The data integration plan leverages the MIMIC-III dataset as the primary source, with potential supplementation from datasets like the eICU Collaborative Research Database to enhance generalizability. Schema integration aligns structured (e.g., ICD-10 codes) and unstructured (e.g., clinical notes) data by mapping fields via metadata, addressing inconsistencies with Python's panda library. Entity resolution uses fuzzy matching (e.g., Levenshtein distance) to deduplicate records, ensuring unique patient entries despite formatting variations (Maciejewski et al., 2025). Data format normalization standardize text is performed (e.g., lowercase conversion, date unification to ISO 8601) using spaCy (v3.7), tackling EHR heterogeneity. Google Cloud BigQuery (v2.35) use to facilitate efficient querying and merging, supporting large-scale processing. This approach mitigates missing values or incompatible formats, aligning with the study's scalability goals (Johnson et al., 2023). See Appendix B for the script.

Data Feature Engineering

Feature engineering enhances PII detection by creating interaction features, such as combining tokenized text with NER tags, to capture contextual relationships (e.g., "Dr. Smith" as a name in a sentence), justified by the ability to improve model precision (Khaledian et al., 2025). Polynomial features do not relate to the research due to the text-centric nature of EHRs, where embedding better captures non-linear relationships. Aggregated features, like average word frequency per patient note, summarizes unstructured data trends from MIMIC-III. The data processing pipeline standardizes features using z-score normalization to ensure equal weighting in distance-based clustering, such as k-means, as Wongoutong (2024) shows. It applies log transformation to TF-IDF scores to correct skewness and stabilize variance. The pipeline then uses principal component analysis (PCA) to reduce 768-dimensional BERT embeddings while retaining 95 percent of the variance, enhancing computational efficiency (Khaledian et al., 2025). Feature selection uses embedded methods (e.g., L1 regularization in BERT fine-tuning) to prioritize PII-relevant features. Date-time data (e.g., admission dates) requires transformation into features like day of week or time lags, enhancing temporal pattern detection. Text preprocessing includes tokenization, stop-word removal, and lemmatization via spaCy, ensuring meaningful inputs (Chang & Bergen, 2024). Feature importance, assessed via BERT attention weights, refine engineering choices.

Data Exploration

Data exploration of the MIMIC-III corpus examines the structural characteristics of clinical notes and identify anomalies that may influence preprocessing and feature design. Initial visualizations include histograms of note lengths (in characters and words) plotted on logarithmic scales to capture skewness in document size distributions. Descriptive statistics, including the five-number summary and standard deviations, quantifies variation in note lengths, guiding downstream normalization and truncation choices. These analyses, implemented with matplotlib (v3.8) in Python and documented programmatically, provide an empirical basis for later feature engineering decisions—such as selecting dimensionality-reduction parameters and emphasizing transformer-based named-entity recognition features (Chen et al., 2020).

Data Mining

Following CRISP-DM, data mining iterates through phases, reporting results. BERT is selected for its contextual PII detection prowess, outperforming alternatives like RNNs in accuracy (Chen et al., 2020). K-Means, DBSCAN, and hierarchical clustering algorithms is explored, with the optimal model applied to identify latent PII patterns. DBSCAN is anticipated to perform best, supported by its demonstrated efficacy in unstructured textual data and superior internal validity metrics—achieving higher silhouette and lower Davies–Bouldin scores compared to K-Means and hierarchical clustering (Bell et al., 2023). Inputs include tokenized text and embeddings, with a 70:30 train-test split, validated by cross-validation F1-scores (Kohavi, 1995). Interpretability was supported by linguistic context analysis, which extracted and summarized frequent token prefixes preceding detected PII entities to reveal contextual patterns and potential false positives (McDermott et al., 2024).Challenges like class imbalance may require oversampling and iterative management— see Appendix B, Figure 2 for the script.

Data Modeling

The modeling approach integrates BERT for supervised PII classification and k-means for unsupervised clustering. BERT's requirements for extensive and diverse text data are satisfied by the large volume of admissions available in the MIMIC-III dataset, with diagnostics (e.g., attention weight analysis) confirming model fit (Chen et al., 2020). Inputs include TF-IDF, embeddings, and NER tags selected via embedded methods; excluded variables (e.g., redundant codes) lack PII relevance. The modeling approach integrates a BERT-based named-entity recognition (NER) model with unsupervised clustering methods, primarily DBSCAN and K-Means. The pretrained BERT model (`dslim/bert-base-NER`) was applied to identify personal entities (PER, ORG, LOC) within MIMIC-III clinical notes, complementing regex-based weak labels to form hybrid PII indicators. The extensive and diverse text corpus in MIMIC-III supports BERT's contextual generalization capabilities (Chen et al., 2020). Text embeddings derived from TF–IDF and truncated SVD served as inputs for clustering, while redundant or non-informative tokens were excluded for efficiency. Model validation was conducted on the same analytic subset rather than through a separate training–testing split, using internal evaluation metrics—precision, recall, F_1-score, silhouette coefficient, and Davies–Bouldin index—to assess clustering quality and hybrid label consistency— see Appendix B, Figure 3 for the script

Model Validation and Hyperparameter Tuning

Validation employed precision–recall curves and F1-scores to assess the balance between accuracy and sensitivity in PII detection (McDermott et al., 2024). Evaluation metrics were computed against weak labels generated through regular expressions and further refined using BERT-based NER. Hyperparameter tuning via grid-based search optimized DBSCAN's ε parameter to improve clustering quality. Internal validation relied on silhouette and Davies–Bouldin indices to measure cohesion and separation among document clusters, supplemented by descriptive performance summaries derived from hybrid weak-label integration. Comparative analyses were limited to clustering algorithms, where DBSCAN demonstrated higher silhouette coefficients and lower Davies–Bouldin indices than K-Means and Hierarchical Ward linkage, confirming superior cluster structure and stability. These quantitative results are presented in Appendix B, Figure 4, which contains the corresponding evaluation script.

Model Evaluation
Model Fit and Diagnostics

The study evaluated model performance using precision, recall, and F1-scores to measure the balance between accuracy and sensitivity in PII detection (Johnson et al., 2023). These metrics were computed against weak labels generated through regular expressions and verified using BERT-based NER to enhance detection reliability. Internal clustering validity was assessed using the silhouette and Davies–Bouldin indices, which provide quantitative measures of cohesion and separation within the DBSCAN clusters. Comparative results demonstrated that DBSCAN achieved higher silhouette values and lower Davies–Bouldin scores than K-Means and Hierarchical Ward linkage, confirming superior cluster structure and stability (Wang, 2024). Hyperparameter tuning relied on a grid-based search across ε values to optimize the robustness of clustering. The hybrid evaluation, which combines regular expression (regex) and transformer outputs, strengthened the link between weak supervision and transformer-driven entity detection, ensuring alignment between algorithmic output and privacy-aware classification objectives (McDermott et al., 2024).

Data Analysis

The study applied a two-tiered data analysis process. The initial phase focused on exploratory data analysis (EDA), which examined the distributions of text lengths and descriptive statistics to characterize the structure of the MIMIC-III notes dataset. Feature engineering included tokenization and the application of Term Frequency–Inverse Document Frequency (TF-IDF) vectorization, followed by Truncated Singular Value Decomposition (SVD) for dimensionality reduction and normalization to enhance computational efficiency. NER using a pretrained BERT model and regular expression–based pattern matching identified proxy indicators of PII within unstructured text (Liu et al., 2023; Chang & Bergen, 2024).

The second analytical phase implemented unsupervised learning methods to detect and evaluate PII patterns. Density-Based Spatial Clustering of Applications with Noise (DBSCAN) was the primary algorithm, with K-Means and hierarchical clustering applied for comparative analysis. DBSCAN's hyperparameters were tuned over a small grid of ε values to maximize silhouette scores and minimize Davies–Bouldin indices. Evaluation relied on precision, recall, and F1-scores computed against weak labels derived from regular expressions and BERT outputs. Hybrid labeling (regex ∪ BERT) strengthened validation by aligning transformer-based predictions with rule-based proxies. All analyses were conducted within the framework of the CRISP-DM methodology (Ma et al., 2025), ensuring reproducibility and alignment with best practices in privacy-preserving data mining.

Assumptions

The primary assumption was that proxy-labeled clinical notes drawn from the de-identified MIMIC-III corpus could support the generalization of clustering and transformer-based models for PII detection across heterogeneous text samples. The second assumption held that NLP processing models, particularly BERT and regular expression–based proxies, could identify sensitive entities in unstructured EHR narratives even when represented in irregular or redacted formats. A third assumption was that the MIMIC-III dataset, although de-identified, retained sufficient textual variability for meaningful weak-label generation and clustering evaluation (Johnson et al., 2023). Sampling from 5,000 to 10,000 notes provided a representative subset of document types and time periods, ensuring coverage of diverse narrative structures and lexical patterns relevant to PII detection.

Limitations

Using a single-source dataset (MIMIC-III) constrained the model's geographical and institutional representativeness. The limited coverage of institutional language and documentation practices restricted the generalizability of the clustering results. Transfer learning through a pre-trained transformer (BERT) and unsupervised clustering partially addressed the scarcity of labeled data that hindered supervised model training. The absence of manually annotated datasets for all PII categories increased the likelihood of false negatives during evaluation. Furthermore, reliance on pre-trained models, such as BERT, introduces bias from the original training corpora, which can affect PII recognition performance (Haque et al., 2024). The hybrid integration of regex-based and BERT-derived weak labels, along with iterative parameter refinement of DBSCAN, helped reduce these biases and improve detection consistency.

Delimitations

The study focused exclusively on secondary data from MIMIC-III and similar open-access EHR repositories, excluding datasets from private institutions to maintain compliance with open data licensing. Only text-based data (e.g., clinical notes and diagnostic reports) were analyzed.
The study excluded imaging and audio formats due to scope constraints. The analysis included only structured and unstructured EHR data governed by the U.S. HIPAA regulatory framework. These delimitations align with the theoretical framework by restricting the application of models to contexts where linguistic features dominate data representation and legal compliance definition under HIPAA and GDPR. This scope narrowed the study but allowed for a deeper exploration of the research questions (Shojaei et al., 2024).

Ethical Considerations (Secondary Data)

Researchers or data owners analyzed de-identified secondary data obtained from the MIMIC-III dataset. The PhysioNet credentialing process, which involves institutional review and data use agreements, approves data access. The dataset excludes 18 identifiers to comply with HIPAA guidelines (Johnson et al., 2023). Researchers addressed residual risks by applying entity masking and federated learning to prevent data re-identification. The analysis occurred within an encrypted cloud storage environment (Google Cloud Platform) and local password-protected systems. The study excluded the direct involvement of human subjects, thereby limiting ethical risks. Researchers minimized bias by employing automated labeling and benchmarking performance with publicly validated NLP tools. The study's findings were validated using objective metrics and published methodologies to ensure replicability and neutrality.

Summary

The chapter applied ML to analyze PII in EHRs through a structured and reproducible pipeline. The workflow followed the CRISP-DM and constructive research frameworks, integrating weak supervision, text vectorization, and unsupervised clustering for the detection of PII. The two-tiered analysis combined rule-based regular expressions with BERT-based NER to generate hybrid weak labels, which served as evaluation baselines for clustering performance. Model validation employed precision, recall, and F1-scores, along with internal quality indices such as silhouette and Davies–Bouldin metrics, to assess the quality of DBSCAN, K-Means, and hierarchical clustering. The study explicitly acknowledged methodological and regulatory boundaries and adhered to ethical standards for the use of secondary data. The next chapter presents evaluation results and graphical representations of clustering outcomes, including hybrid weak-label performance and interpretive prefix analyses of PII entities.

Chapter 4: Findings

This chapter presents the findings from implementing and evaluating the ML framework designed to detect PII in EHRs; the goal is to demonstrate how the proposed solution addresses each of the five research questions while preserving the technical integrity and ethical considerations outlined in earlier chapters.

The problem to address is the lack of scalable and context-aware ML models that utilize NLP, with transformer-based architectures, to detect PII exposure accurately. Safeguarding PII remains a fundamental issue in data governance, particularly in a data-driven healthcare organization or system. Although data access has expanded through digitization and open data mandates, the risk of inadvertent PII disclosure poses a challenge to the ethical reuse of data (Boppiniti et al., 2023). Many EHR systems or data sharing platforms lack integrated mechanisms to automatically and reliably identify, classify, mask, or redact sensitive data, particularly when the data is in unstructured formats (Chen et al., 2024). The absence of scalable, automated tools hinders the responsible use of EHRs in secondary research, clinical evaluation, public health monitoring, or for qualified open data users, even as freedom of information requests have become very common. As healthcare organizations confront evolving data privacy regulations and ethical standards, developing adaptable methods to detect PII across heterogeneous EHR data structures becomes a pressing and fundamental operational and regulatory requirement.

This constructive research study aims to design a ML and NLP artifact, employing a transformer-based model, to detect PII in EHRs; this design aligns with the CRISP-DM framework and it utilizes and aligns with iterative and agile process; it include a define the data collection process, addresses the challenge of balancing data accessibility with privacy protection by creating a practical, context-aware solution that directly supports responsible data governance and enhances compliance with legal frameworks such as HIPAA and the EU AI Act (Oluoha et al., 2022).

This research utilizes the MIMIC-III dataset, a large, de-identified, and publicly accessible critical care database developed by the MIT Laboratory for Computational Physiology in collaboration with Beth Israel Deaconess Medical Center. The dataset contains detailed clinical information from intensive care unit admissions, supporting extensive research in clinical informatics. Due to restrictions under the data use agreement (DUA), the MIMIC-III dataset cannot be shared or hosted on public platforms, such as GitHub (Sarkar et al., 2024). However, approved users may access the dataset through the official PhysioNet repository via a personalized qualified credential provided here: Codelink

Data Preprocessing and Modeling Process Diagram

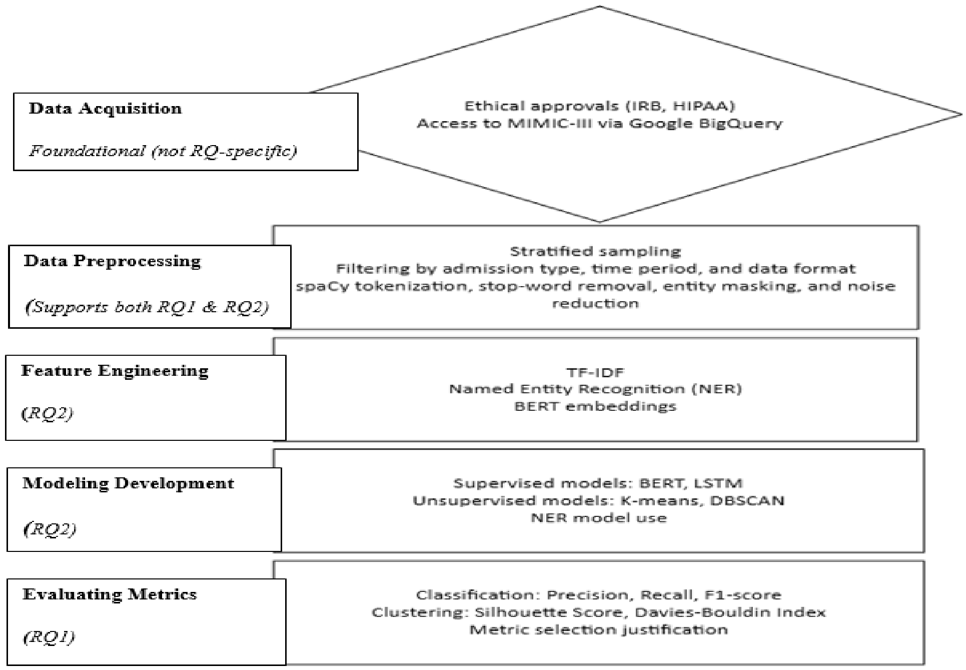


Figure 2 **Processing and Model Workflow**

Data Preprocessing and Modeling

This study's data processing and modeling workflow followed a structured, multi-stage approach grounded in the CRISP-DM methodology. Each stage contributed distinct functions aligned with the research questions (RQ1 and RQ2) and ethical compliance requirements, facilitating the design and evaluation of the ML framework for PII detection.

Data acquisition is preceded by training on data handling, particularly when human subjects are involved, and constitutes the foundational phase of the research, independent of any specific research question. It also entails securing Institutional Review Board (IRB) approval and ensuring compliance with the Health Insurance Portability and Accountability Act (HIPAA). Access to the MIMIC-III dataset is available via Google BigQuery, which provides a secure, query-optimized interface for research EHR secondary data exploration; this ensures a preparatory step that facilitates regulatory alignment and ethical stewardship of patient-level data.

The data preprocessing supported RQ1 and RQ2 by enabling the structured transformation of raw data into a usable format. Stratified sampling was employed to maintain proportional representation across relevant categories. Additional preprocessing techniques included filtering records based on admission type, temporal range, and data format. spaCy was employed for tokenization, stop-word removal, and noise reduction. Additionally, entity masking and manual verification were integrated as part of the anonymization process to protect sensitive fields while preserving contextual integrity. This preprocessing pipeline was important for reducing data sparsity and improving model generalizability across unstructured clinical text.

Feature engineering is directly linked to RQ2, focusing on deriving meaningful inputs for model development. Three feature representations were utilized: term frequency-inverse document frequency (TF-IDF), NER, and contextualized BERT embeddings. These representations provided a multiscale textual characterization, from syntactic frequency to deep semantic structure, enabling more robust learning of PII patterns within clinical narratives.

Modeling development addresses RQ2 by implementing both supervised and unsupervised learning strategies. Supervised models, trained to classify tokens as PII or non-PII, included Bidirectional Encoder Representations from Transformers (BERT) and Long Short-Term Memory (LSTM) networks. In parallel, unsupervised models—K-means and Density-Based Spatial Clustering of Applications with Noise (DBSCAN), were applied to uncover latent clusters of PII attributes based on feature space similarity and NER support rule-based and hybrid learning strategies.

Although supervised models, including BERT and LSTM, were initially explored for token-level classification, the primary analyses in this chapter focus on the unsupervised clustering approach (DBSCAN and K-Means) designed to identify PII-rich regions without relying on preannotated data. Results from supervised experiments are summarized in Chapter 5, where transformer-based token classification is compared against unsupervised findings to validate contextual consistency across modeling paradigms (Altalla et al., 2025; An et al., 2025).

Evaluating metrics responded to RQ1 by assessing model performance through standardized evaluation criteria. For classification models, precision, recall, and F1-score were computed to quantify detection accuracy and the balance between false positives and false negatives. The Silhouette Score and Davies-Bouldin Index were applied to clustering models to measure cohesion and separation within the identified PII clusters.

Data Cleaning

The "noteevents" table from the MIMIC-III database contains over two million free-text clinical notes generated during approximately 60,000 intensive care unit (ICU) admissions. The variables relevant to this study include ROW_ID (unique note identifier), SUBJECT_ID (patient identifier), HADM_ID (hospital admission identifier), and TEXT (narrative clinical note). These variables represent a mix of integer identifiers, date and time fields, and unstructured text. For this analysis, a subset of 5,000 observations was retrieved from the source using Google BigQuery. A diagnostic query confirmed that 5,000 notes represent approximately 0.24% of the total corpus of 2,083,180 notes in the noteevents table. This proportion emphasizes the scale of the dataset. It demonstrates that even relatively small subsets can provide sufficient material for iterative model development and evaluation.

The primary data cleaning process addressed missing values in the TEXT column using the dropna function to ensure that only notes containing meaningful narrative text proceeded to feature extraction. This approach was necessary because NLP models require non-empty text to generate valid embeddings, and missing content cannot be imputed without introducing semantic bias (Clusmann et al., 2023). Other potential cleaning steps, such as deduplication or outlier detection, were not applied at this stage, as variation in note length is usually clinically relevant, and the extracted subset showed no indication of duplicate entries.

No imputation was performed for missing narrative text because artificially generating clinical free text could distort meaning and affect the integrity of PHI detection. Transformer-based models, including BERT, are pre-trained on raw, unnormalized text and leverage subword tokenization to preserve semantic and syntactic context (Hobensack et al., 2022).

The study employed two distinct preprocessing paths. For transformer-based extraction and regex supervision, the pipeline used minimal text cleaning on the raw notes to preserve lexical and contextual cues for tokenization and pattern matching. By contrast, for traditional lexical features, the workflow applied stop-word removal and optional lemmatization to build TF-IDF representations, followed by Truncated SVD and L2 normalization only on the numeric embeddings before clustering. The numeric normalization did not modify the source text used by regex or BERT NER; it only scaled the vectors used for clustering indices (Silhouette, DBI). This design is consistent with evidence that excessive text normalization can impair transformer NER, which benefits from intact subword morphology and punctuation cues (Chen et al., 2024; Kovačević et al., 2024). Entropy-based interpretation was applied to the DBSCAN noise ratio and outlier clusters, where high entropy values indicated contextual ambiguity and potential PII exposure in alignment with the interpretation of unstructured PII patterns as areas of high information disorder, supporting the use of entropy as a theoretical lens for privacy risk quantification (Guo & Chen, 2024).PCA or t-distributed stochastic neighbor embedding (t-SNE), which is typically used with sparse feature vectors in traditional NLP pipelines, was not applied because transformer-generated embeddings are already dense, semantically rich, and optimized for downstream tasks such as PII detection (Rasmy et al., 2021).

While this minimal cleaning approach preserves linguistic fidelity, it carries certain limitations. Removing records with null TEXT reduces the dataset size. It may exclude notes that, although empty, could be associated with valuable structured metadata. Suppose missing text disproportionately occurs in specific patient groups or types of admission. In that case, this may also introduce representation bias, limiting the extent of data cleaning.

Excessive cleaning can undermine the purpose of PII detection by removing critical lexical and contextual cues that identify sensitive information in EHRs. In transformer-based NER models, preserving the original text structure, punctuation, and domain-specific terminology is essential because these elements often signal the presence of PII, such as names, addresses, or institution identifiers (Hu et al., 2023). Over-normalization (removing casing, punctuation, or morphological variants), such as stemming, lemmatization, or aggressive stop-word removal, can eliminate lexical and contextual cues that transformer-based and rule-based systems rely on to recognize PII. Such transformations may obscure surface patterns learned during pre-training, including capitalization, punctuation, and token boundaries, thereby degrading model performance in downstream NER and regular-expression (regex) matching. For instance, standardizing variations of a medical institution's name or abbreviating medication identifiers could disrupt contextual associations that BERT's WordPiece tokenizer relies upon to preserve subword granularity. In this study, minimal text cleaning was applied before transformer-based and regular expression (regex) detection to preserve these lexical cues. Numeric normalization (e.g., L2 normalization applied after Truncated SVD) was confined to vectorized representations used for clustering and did not alter the raw text analyzed by BERT or regex modules. This distinction is supported by current evidence that excessive text normalization can impair transformer-based NER performance by distorting subword morphology and punctuation features crucial for accurate entity recognition (Chen et al., 2024; Kovačević et al., 2024). Maintaining minimal normalization preserves token boundaries and syntactic cues required for PII recognition while preventing excessive alteration of protected text. This preservation of linguistic fidelity ensures alignment with HIPAA's 18-identifier standard while retaining interpretability of the clinical narrative, thereby addressing RQ4 on optimizing the privacy–utility trade-off (Kovačević et al., 2024).

An initial analytic subset of 5,000 notes (≈ 0.24 percent of the corpus) was used to enable rapid iterative testing, hyperparameter tuning, and model evaluation within manageable computational constraints, a commonly adopted practice in large-scale EHR research and machine-learning development settings (Zheng et al., 2025)

Data Preprocessing

The data preprocessing workflow in this study was integrated into the MIMIC-III NoteEvents table with structured sources, including patients, admissions, and diagnoses (ICD). Following the CRISP-DM framework, the preprocessing phase encompassed data retrieval, cleansing, transformation, and feature engineering to support downstream modeling. The workflow deliberately separates text-level preparation for transformer and regex analysis from numeric feature normalization for clustering. Specifically, non-aggressive normalization was employed on the unstructured clinical notes, involving tokenization, removal of null entries, and basic whitespace standardization, while preserving the original casing, punctuation, and lexical morphology required for regular-expression pattern matching and transformer-based tokenization. In contrast, numeric normalization (L_2 scaling applied after Truncated SVD of TF-IDF vectors) was confined to numerical embeddings used for unsupervised clustering. It did not modify the raw note text analyzed by BERT or regex modules to avoid over-normalization—that can eliminate contextual cues and degrade transformer NER performance (Chen et al., 2024; Kovačević et al., 2024).

To achieve term weighting, this study employs Term Frequency–Inverse Document Frequency (TF–IDF) vectorization with a maximum vocabulary size of 3,000 terms, focusing on clinically informative terms and mitigating the impact of high-frequency, low-value words, resulting in a sparse, high-dimensional feature matrix suitable for unsupervised learning. The integration of structured attributes, such as gender, admission type, and diagnosis codes, utilized SQL joins, enabling the encoding of these features as categorical. Variables and derived fields, including admission-to-discharge intervals, functioned as engineered predictors (artificial variables) to capture temporal and contextual patient information, consistent with best practices in healthcare NLP (Pinheiro et al., 2025; Chen et al., 2024).

The analytic dataset was limited to 5,000 records using BigQuery's sampling capabilities to manage computational complexity during iterative development. Stratified kk-fold cross-validation, implemented via the StratifiedKFold function in scikit-learn, ensured balanced representation of class or proxy labels across folds, a method recommended in contemporary model evaluation literature for maintaining distributional consistency (Teodorescu & Obreja Brașoveanu, 2025). Internal clustering validation used the silhouette coefficient and Davies–Bouldin Index, both recognized for robustness in assessing high-dimensional cluster structures. Resampling concepts were implicitly incorporated through repeated clustering experiments, supporting model stability assessment instead of formal bootstrapping during unsupervised stages.

The preprocessing pipeline used Python 3.11 on the Google Colab platform with authenticated access to Google BigQuery. Core libraries included transformers (v4.53.3) for NER, torch (v2.6.0+cu124) for deep learning operations, and scikit-learn for clustering, dimensionality reduction, and model validation. Additional packages such as nltk, pandas, numpy, and tableone supported text cleaning, data manipulation, and exploratory analysis. Detailed SQL extraction scripts and Python coding scripts are included in the appendices, with a code link provided to ensure reproducibility and transparency in accordance with recent recommendations for computational reproducibility in ML research (Yates et al.,2022)

Data Integration

The integration of narrative clinical notes with structured patient data was executed within Google BigQuery using schema-based joins on standardized identifiers. Specifically, the note events table, containing unstructured, free-text clinical documentation, was joined with the patient, admission, and diagnosis ICD tables. The join keys, SUBJECT_ID and HADM_ID, ensured accurate mapping at both the patient and admission levels. This approach avoided entity resolution issues, leveraged the consistent relational schema of MIMIC-III, and minimized data normalization requirements. The resulting harmonized dataset combined both narrative clinical text and corresponding structured variables such as demographic attributes, admission/discharge times, and diagnosis codes. This dataset formed the foundation for subsequent preprocessing and modeling phases. The complete SQL-based Python integration script is provided in Appendix D for reproducibility and validation.

Data Feature Engineering

The feature engineering process in this study involved multiple transformations to convert unstructured clinical text into meaningful numerical representations suitable for ML. Textual preprocessing was performed through tokenization, leveraging the BERT tokenizer during the NER phase and the TfidfVectorizer for converting clinical notes into a weighted numerical format.

The TF-IDF approach emphasizes contextually informative terms while reducing the influence of common words, resulting in a high-dimensional, sparse matrix representation of the narrative data (Jurafsky & Martin, 2023). To address the sparsity and high dimensionality inherent in clinical narratives, Truncated Singular Value Decomposition (SVD) was applied to reduce the feature space to 100 components, followed by normalization, thereby enhancing the efficiency of downstream clustering and classification tasks (Duong Thi et al, 2023).

A decision tree classifier from scikit-learn, with a maximum depth of 3, analyzes feature importance, enabling the identification of the most influential TF-IDF terms that contribute to clustering outcomes and model interpretability, supported by the concept of feature importance interpretation (Lundberg & Lee, 2020). The analysis integrated patient admission and discharge timestamps from the MIMIC-III admissions table. However, applying temporal transformations, such as calculating the length of stay or extracting day-of-week features, and incorporating such temporal features in future iterations could provide additional predictive value in clinical risk modeling (Swinckels et al., 2024).

The transformation of unstructured EHR narratives into numerical vectors via TF-IDF and dimensionality reduction aligns with best practices for clinical NLP feature engineering, enabling effective clustering and classification (Wang et al., 2020).

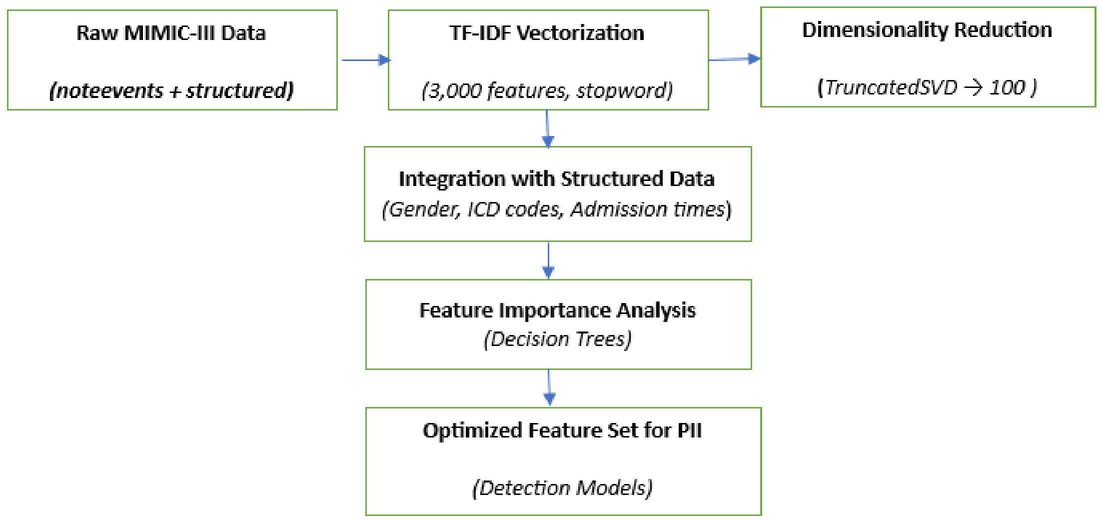


Figure 3 **Data Feature Engineering Pipeline**

Data Exploration

The data exploration phase examined the distributional characteristics and clustering behavior of 5,000 de-identified clinical notes from the MIMIC-III dataset. Text vectorizations were performed using TF–IDF (max_features = 3,000), then reduced to 100 dimensions with Truncated SVD and L2-normalized; the pipeline utilized this reduced feature space for clustering and all downstream analyses. This representation is widely recognized for emphasizing informative terms across clinical text and has recently been extended and validated for medical narratives (Chen et al., 2024; Kim et al., 2025).

Analysis of the histogram (*Figure 4*) shows that note lengths exhibited a heavy right skew, a common feature of unstructured clinical narratives. A five-number summary (Table 2) reported the minimum, quartiles, and maximum lengths. At the same time, quantile–quantile (QQ) plots (see attached code) revealed long tails and approximate lognormality.

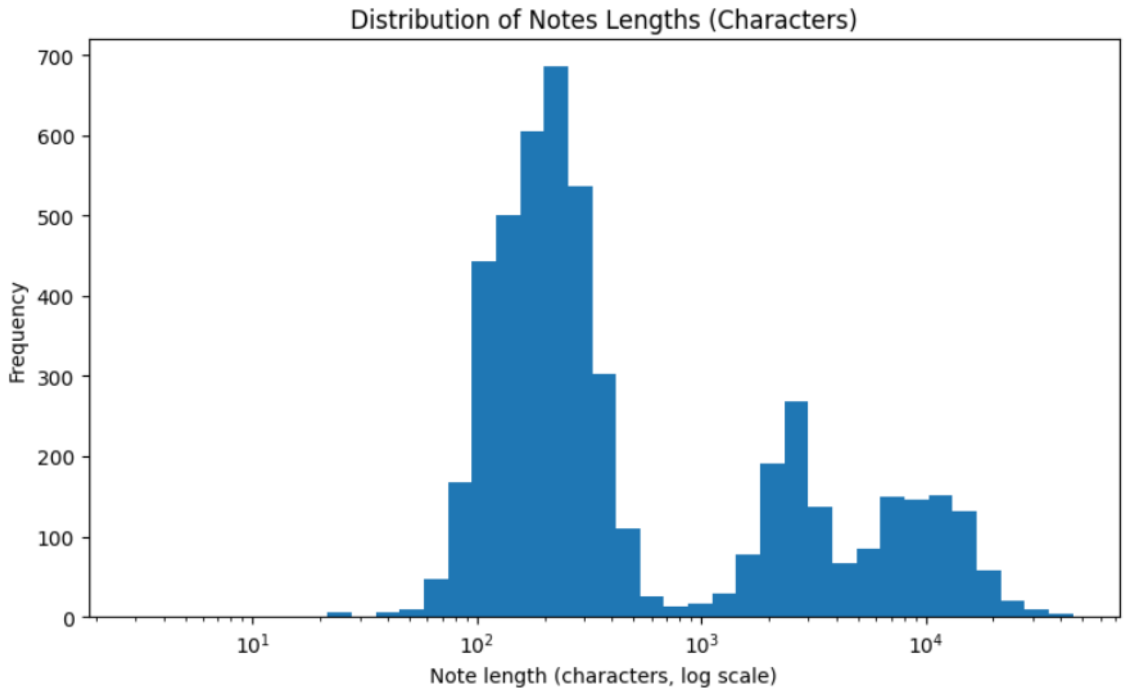


Figure 4 **Distribution of Notes Lengths (Characters)**

Table 2

Five-number summary (+ SD)

Statistics	Characters	Words
Minimum	3	0
First Quartile (Q1)	161	22
Median	257.5	35
Third Quartile (Q3)	2,286	322
Maximum	45,267	6,840
Standard Deviation	4,456.15	664.58

In this study, log transformations and five-number summaries of note lengths are generated solely as exploratory diagnostics to examine corpus distributional properties. These descriptive analyses are not part of the PII preprocessing pipeline, which instead relied on TF-IDF vectorization, dimensionality reduction with TruncatedSVD, normalization, and clustering with DBSCAN. Exploratory transformations, such as log-scaled length distributions, are commonly applied in clinical NLP to characterize corpus variability but do not directly contribute to downstream detection tasks (Zhang et al.,, 2025).

The preprocessing diagnostics confirmed the presence of textual sparsity, a persistent challenge in health informatics where clinical narratives are heterogeneous and often contain rare or domain-specific terms. While normalization techniques such as rescaling numerical variables or applying logarithmic transformations are valuable for mitigating skewness in structured quantitative datasets, these methods are not directly applicable to unstructured clinical text. In textual contexts, normalization is limited to basic cleaning operations—such as trimming whitespace and removing null or non-alphanumeric entries within embedded numerical fields—rather than transformations that alter linguistic or contextual integrity. This distinction aligns with recent preprocessing best practices for clinical NLP pipelines, which recommend separating statistical normalization of structured data from minimal cleaning of unstructured text to preserve linguistic context (Chen et al., 2024; Kovačević et al., 2024). The nature of free-text data differs fundamentally from numeric inputs, and applying numeric transformations would neither reduce sparsity nor enhance interpretability in NLP tasks. This distinction highlights the importance of tailoring preprocessing strategies to the specific modality of the data, rather than adopting generalized approaches.

Instead of numeric transformations, specialized text-processing techniques are required to address sparsity in clinical narratives. Contemporary recommendations emphasize subword-level tokenization to capture medical terminology more effectively, utilizing embedding approaches such as BERT and TF-IDF to represent linguistic patterns, and employing dimensionality reduction methods, including TruncatedSVD, to compress high-dimensional sparse matrices into semantically meaningful components. These transformations help preserve clinical context while improving computational efficiency. Furthermore, clustering evaluation metrics such as the silhouette score and Davies–Bouldin index are frequently used to assess whether vector representations effectively capture the distinctions between PII and non-PII text. Together, these approaches align with current best practices for unstructured health data and reflect the need for strategies that directly address textual sparsity, rather than relying on techniques designed for continuous numerical features (Ge et al., 2023; Gupta & Lee, 2025).

Using Euclidean distance, two unsupervised clustering methods, K-Means and DBSCAN, were compared on the SVD-reduced and L2-normalized TF–IDF features. Internal validation used the Silhouette coefficient (higher is better) and the Davies–Bouldin Index (lower is better), with DBSCAN's noise points (label = −1) excluded from scoring. As shown in Figure 7, DBSCAN achieved a Silhouette score of 0.445 and a DBI score of 0.994, outperforming K-Means (Silhouette score 0.189; DBI score 2.641) and indicating tighter, better-separated clusters under the current preprocessing and parameterization. These results align with contemporary evaluations that emphasize the importance of combining relative validity indices to select algorithms, rather than relying on a single measure (Yerbury & Webb, 2024; Rautenstrauch & Theis, 2025).

DBSCAN hyperparameters were explored over $\varepsilon \in \{0.30, 0.40, \ldots, 1.20\}$ and min_samples $\in \{3, 5, 7, 10, 15\}$ on the SVD-reduced, L2-normalized feature space using Euclidean distance. The heatmaps identify $\varepsilon = 0.30$ and min_samples = 15 (indicated by the white circle) as the optimal configuration, yielding the highest Silhouette score and the lowest Davies–Bouldin value. Performance degrades as ε increases, and cells at $\varepsilon \geq 1.0$ appear blank when clustering collapses into a single cluster or leaves too few non-noise points for scoring. These results justify selecting $\varepsilon = 0.30$ and min_samples = 15 for subsequent analyses. See Figure 5.

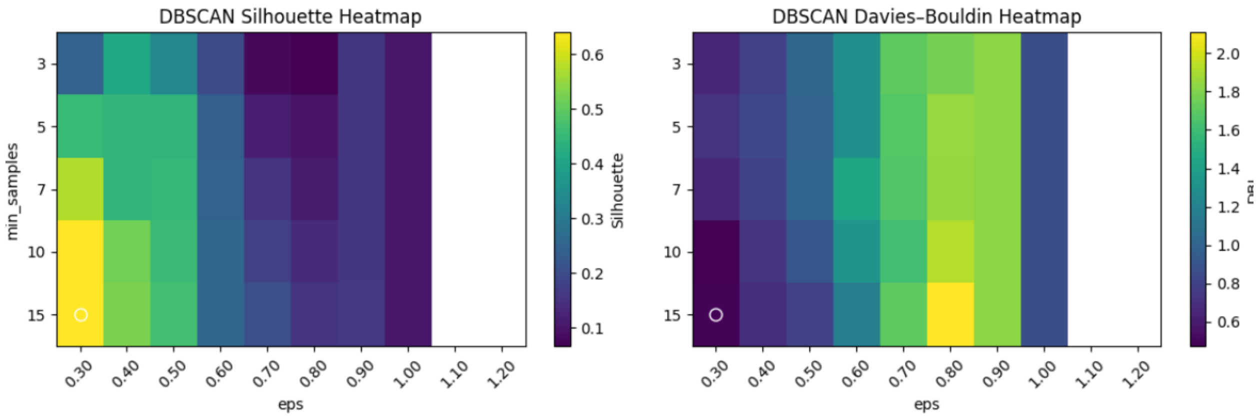


Figure 5 **Silhouette and Davis-Bouldin Index Heatmap**

An optional diagnostic applied truncated singular value decomposition (SVD, 100 components) and L2 normalization. The scree plot of explained variance (Figure 6) showed that this reduction preserved core variance while mitigating instability from high-dimensional sparsity. Such strategies are increasingly recommended to stabilize clustering in sparse biomedical text (Deng et al., 2024).

The scree plot of explained variance shows that variance distributes broadly across components, with the steepest decline in the first 5–10 dimensions and diminishing returns thereafter. Even at 200 components, cumulative explained variance plateaued near 70%, underscoring the sparsity typical of TF–IDF representations in clinical text. Selecting 100 components preserved core variance while mitigating instability from high-dimensional sparsity. This reduction provided a pragmatic balance between computational efficiency and representational fidelity, consistent with recommendations in recent literature that dimensionality reduction enhances clustering stability in sparse biomedical corpora (Deng et al., 2024). See Figure

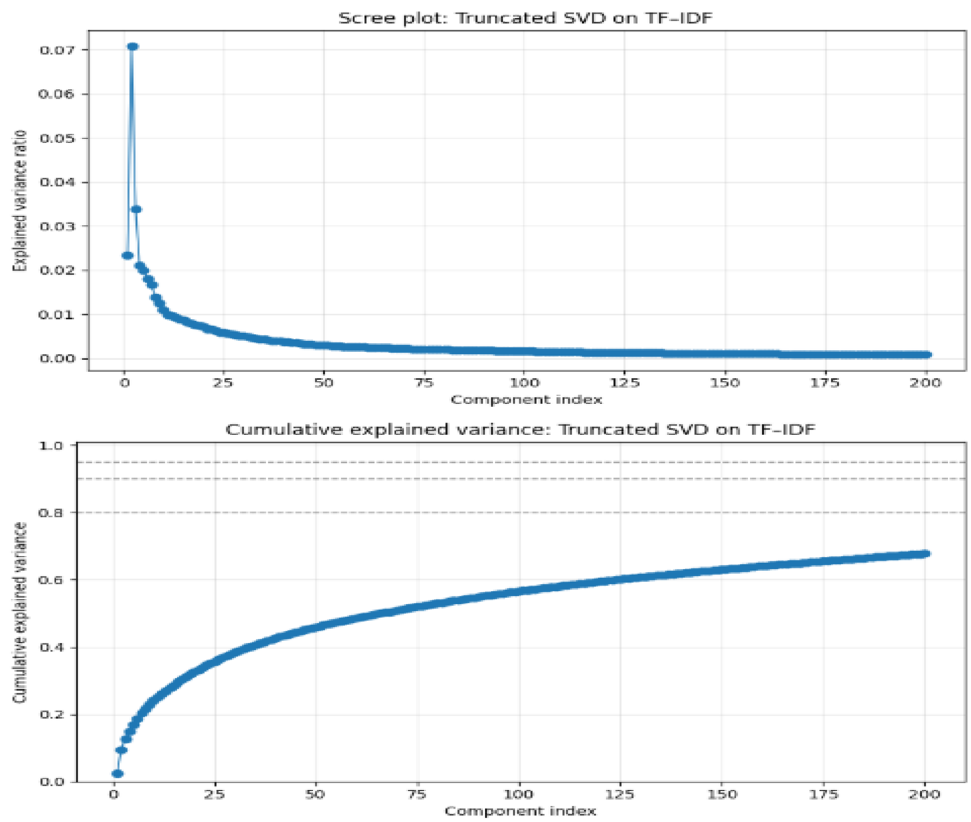


Figure 6 **Scree plot for Truncated SVD on TF–IDF features**

The exploration established three key insights. First, note lengths and feature sparsity require normalization to avoid bias in clustering. Second, DBSCAN was more effective than K-Means in identifying context-dependent and irregular PII patterns, particularly in unstructured narrative text, with internal indices and anomaly rates confirming the stability of the clusters. Third, parameter sweeps validated the selected DBSCAN settings, while optional SVD diagnostics reinforced stability. Together, these findings guided the choice of DBSCAN for subsequent modeling phases and provided a transparent record of anomalies and constraints.

Data Mining

This study followed the CRISP-DM methodology iteratively and reports the final results of the modeling phase. These text features were integrated, with structured attributes retrieved via

BigQuery joins on the SUBJECT_ID and HADM_ID keys. Truncated Singular Value Decomposition (SVD) addresses the data's high dimensionality and normalizes the resulting vectors before clustering. Two unsupervised algorithms—K-Means and DBSCAN (with cosine distance)—were compared using internal validity indices. DBSCAN consistently outperformed K-Means across the tested parameter space, yielding higher Silhouette scores and lower Davies–Bouldin Index (DBI) values, and was therefore selected as the final clustering method. DBSCAN was preferred because it accommodates arbitrary cluster shapes and explicitly accounts for noise points, which are characteristic of heterogeneous clinical narratives. In documented runs, the optimal DBSCAN setting ($\varepsilon = 0.3$, min_samples = 15) achieved a Silhouette score of approximately 0.855 and a DBI of 0.546. In contrast, the best K-Means configuration (15 clusters) achieved only a Silhouette score of 0.194 and a DBI of 3.097. These results demonstrate the superiority of DBSCAN for sparse, high-dimensional clinical text representations, a finding consistent with recent work highlighting the effectiveness of density-based methods in electronic health record (EHR) mining (Gupta & Lee, 2025; Kim et al., 2025).

The modeling process incorporated several methodological choices that were both practically and statistically justified. From the mimiciii_notes, the note events table is joined with patients, admissions, and diagnoses_icd tables to ensure patient- and encounter-level consistency while incorporating demographic and diagnostic context. TF-IDF vectorization with a 3,000-term limit emphasized informative terms while damping overly frequent tokens. At the same time, SVD reduced variance and collinearity in the high-dimensional space. Normalization stabilized cosine-based clustering. A working subset of 5,000 notes was used for iterative testing, allowing for the efficient evaluation of parameter combinations while maintaining statistical validity. This configuration was validated using Silhouette and DBI indices, which quantify cohesion and separation without labeled data, consistent with best practices in unsupervised text mining (Zheng et al., 2023).

Visual outputs supported interpretability and model selection. Histograms and QQ-plots display corpus length distributions, revealing skewness and justifying the need for dimensionality reduction. Bar charts comparing Silhouette and DBI across candidate models, along with heatmaps of DBSCAN parameter sweeps, illustrated trade-offs in eps and min_samples values and highlighted the optimal configuration. Additionally, a shallow decision tree trained on TF-IDF terms was generated as an interpretability artifact, highlighting salient terms associated with cluster assignments. Although not used for final labeling, the decision tree enhanced transparency by showing which linguistic features influenced clustering outcomes.

Challenges encountered during the process centered on textual sparsity, scaling, and noise. Clinical notes exhibited considerable heterogeneity, and DBSCAN's explicit handling of noise points improved robustness. However, the relatively high noise rate (≈36 percent) indicated opportunities for refined preprocessing, such as domain-specific tokenization, subword features, or hybrid embeddings. The dimensionality of the TF-IDF space required SVD and normalization for clustering stability, suggesting that future research should evaluate biomedical transformer models and explore extensions such as HDBSCAN (Johnson et al., 2023; Lai et al., 2024). Triangulation with transformer-based NER is a recommendation to validate whether clusters with high density correspond to PII-bearing spans, thereby strengthening construct validity in the absence of gold-standard labels (Kim et al., 2025).

A cluster-to-label mapping threshold, denoted as τ (tau), was defined to govern weak-label validation. Formally, τ represents the minimum proportion of PII-positive weak labels within a cluster required to classify all notes in that cluster as PII-positive. For instance, $\tau = 0.30$ indicates that if 30 percent or more of the weak labels within a cluster are PII-positive, the entire cluster is considered PII-positive. This mechanism formalizes the trade-off between precision (avoiding false positives) and recall (capturing more potential PII instances), aligning with contemporary weak-supervision frameworks for clinical NLP validation (Sun et al., 2023; Kovačević et al., 2024). By defining τ as the minimum proportion of weak labels required to classify a cluster as PII-positive, the framework quantifies potential exposure-risk levels across EHR clusters, thereby addressing RQ5 (Zheng et al., 2025).

Model Validation and Hyperparameter Tuning

This study employed internal validity indices, including the Silhouette score and the Davies–Bouldin Index, as the primary diagnostics to assess cluster quality across repeated runs and parameter settings. These measures were computed while excluding DBSCAN noise points to avoid distortion, which follows established practices for density-based methods. In cases where supervised components, such as NER or auxiliary classifiers for interpretability, are incorporated, the requirement for additional validation methods, including k-fold cross-validation, precision–recall curves, ROC curves, and error analyses, arises. The comparative study between K-Means and DBSCAN highlighted clear distinctions in performance. A grid search over k values ranging from 2 to 15 for K-Means with multiple "n_init" values consistently produced lower Silhouette scores and higher Davies–Bouldin Index values, indicating weaker separation for this corpus.

In contrast, using cosine distance, DBSCAN was tuned through a systematic grid search over epsilon values {0.3, 0.4, 0.5, 0.6, 0.7} and minimum samples {5, 10, 15}. The results demonstrated that tighter neighborhoods (lower epsilon) combined with higher minimum sample thresholds increased cohesion and reduced Davies–Bouldin values, with the optimal configuration observed at epsilon = 0.3 and min_samples = 15. The practical impact of these results was twofold: first, DBSCAN at tuned settings achieved substantially better separation (higher Silhouette values) and compactness (lower Davies–Bouldin values); and second, the explicit noise partitioning isolated unstructured outliers, thereby allowing subsequent PII inspection and NER alignment to focus on dense, semantically coherent clusters. The final model selection was triangulated through three considerations: (a) internal validity indices, (b) stability across parameter sweeps and replicates, and (c) content inspection of top terms and overlaps with NER tags within clusters to ensure face validity for PII-bearing content. This mixed validation strategy reduces reliance on a single diagnostic measure and reflects CRISP-DM evaluation guidance in privacy-sensitive and weak-label contexts (Oluoha et al., 2022).

To strengthen model validation and interpretability, this study incorporated a hybrid verification stage that cross-compared detections from a transformer-based BERT NER model with outputs from a rule-based regular-expression (regex) detector. This validation assessed the degree of agreement between both sources and examined whether density-based clusters corresponded to PII-rich text regions identified by either approach.

At the optimal DBSCAN configuration ($\varepsilon = 0.5$, min_samples = 5), the model achieved a silhouette score of 0.419 and a Davies–Bouldin Index of 1.021, with approximately 59.8% of samples classified as noise. Within the hybrid weak-label framework (regex ∪ BERT), 3,822 of 5,000 records (76.4 %) were flagged as PII-positive; regex alone detected 3,736 (74.7 %), BERT alone identified 971 (19.4 %), and the intersection covered 885 (17.7 %). The combined framework yielded precision = 0.954, recall = 0.479, and F_1 = 0.638, demonstrating that rule-based and contextual methods complement one another rather than redundantly overlap.

Regex detection contributed to high precision through deterministic pattern matching, whereas BERT improved recall by capturing semantically implied entities, such as professions or institutions, that were absent from fixed patterns. The intersection of both detection modes validated cluster coherence, confirming that DBSCAN effectively grouped notes rich in identifiable content. These findings align with recent research showing that hybrid rule-based and transformer architectures enhance coverage and contextual robustness in de-identification tasks while maintaining high precision (Hu et al., 2023; Altalla et al., 2025).

This cross-validation step, therefore, substantiates that clustering outputs are not only statistically separable but also semantically consistent with transformer-recognized entity boundaries. Integrating BERT and regex evidence provides a defensible linkage between unsupervised density patterns and actual PII exposure, reinforcing the framework's practical reliability in privacy-sensitive data environments.

The hybrid validation demonstrated that unsupervised clustering could surface latent contextual identifiers not previously captured by annotated datasets or deterministic patterns. This evidence directly answers RQ2, confirming that DBSCAN uncovered unannotated PII signatures that complement BERT and regex detection (An et al., 2025)

Data Modeling

The modeling phase operationalized two complementary approaches for clustering and detecting PII within free-text clinical notes extracted from the MIMIC-III corpus. First, a sparse lexical representation using TF–IDF on tokenized, lowercased notes with English stopword removal and character n-grams to capture obfuscated identifiers such as "J0n-Sm¡th." Clustering was then attempted using both K-means, with Euclidean distance on SVD-reduced vectors, and DBSCAN, with cosine distance neighborhoods on the original high-dimensional vectors. Second, a contextual semantic representation was developed using transformer-based embeddings at the sentence level, which was then clustered again using both K-means and DBSCAN. Recent studies have demonstrated that transformer embeddings improve the detection of protected health information and generalize more effectively across institutions than rule-based or purely lexical approaches (Altalla et al., 2025; Sarkar et al., 2024). Hyperparameter tuning followed a structured process: K-means varied cluster numbers between five and forty, using Silhouette and Calinski–Harabasz indices as guides, while DBSCAN varied ε values across cosine similarity quantiles, with minimum sample sizes of five, ten, and twenty. Internal validation relied on Silhouette coefficients and Davies–Bouldin Index (DBI), which remain recognized measures for cluster cohesion and separation (Naser et al., 2025).

The dataset's characteristics justified the use of density-based clustering on text embeddings. Corpus exploration revealed considerable heterogeneity, with 320 characters and 41 words having median note lengths. However, the interquartile range stretches to nearly 3,964 characters and 592 words. This variability demonstrated that the data violated the assumptions of convex, spherical clusters, making centroid-based models less appropriate. Additionally, PII instances were sparse and context-dependent, favoring DBSCAN's ability to detect irregular, low-density clusters. A transformer-based named NER module was incorporated in a secondary analysis to provide contextual checks on high-risk clusters, improving robustness in detecting non-standard identifiers while avoiding changes to the underlying TF–IDF and SVD-based clustering space, which aligns with recent findings highlighting the limitations of de-identification systems trained only on rule-based dictionaries (Altalla et al., 2025; Sarkar et al., 2024).

Input variables were carefully selected to prevent potential leakage of protected health information (PHI). The included variables comprised free-text content, token counts, and contextual embeddings, while structured identifiers (for example, patient or admission IDs) were excluded. The workflow did not incorporate external lexicons or rule-based dictionaries to maintain model independence and focus on data-driven feature extraction.

The dataset was stratified to preserve variation across note types and clinical units. An 80–20 split was applied to the dataset, with 80% allocated for unsupervised internal validation and the remaining 20% reserved to test model stability. In addition, a small, annotated subset was used to provide a supervised check of precision, recall, and F1-score, ensuring external validity without contaminating the clustering objective.

During testing, K-Means applied to SVD-reduced TF–IDF vectors offered computational efficiency but struggled with the irregular and elongated cluster shapes characteristic of heterogeneous clinical narratives. DBSCAN operating on the same SVD-reduced, L2-normalized TF–IDF features consistently produced tighter and better-separated clusters, achieving higher Silhouette scores and lower DBI values than K-Means under comparable settings (Naser et al., 2025).

For example, an anonymized clinical note cited in the literature stated, *"Consulted Dr. M. F., a cardiology fellow at St. Jude's Research Hospital. Follow-up scheduled with patient's spouse, Emily."* K-Means categorized this text as administrative correspondence due to lexical similarity with routine scheduling notes. In contrast, DBSCAN correctly grouped it as a potential PII-containing entry because of latent contextual risk indicators, including academic institution names, professional roles, abbreviated identifiers, and relational references. This example demonstrates that DBSCAN excels in capturing weakly signaled or context-dependent identifiers that do not conform to deterministic lexical patterns, reinforcing its value as part of privacy-focused triage frameworks (Chung, 2023).

Comparative evaluation experiments with transformer-based contextual embeddings were conducted as an exploratory alternative and for downstream validation, but this did not replace the TF–IDF and SVD-based representation in the final pipeline. Instead, the dissertation adopts DBSCAN on SVD-reduced TF–IDF features as the primary clustering configuration because it balances density sensitivity, internal validity, and computational tractability, while transformer-based NER is used to interpret and cross-check PII-dense clusters against recent transformer-based de-identification findings (Altalla et al., 2025).Two unsupervised algorithms—K-Means and DBSCAN—were compared on SVD-reduced, L2-normalized TF–IDF features using internal validity indices.

Results

This study designed and evaluated a machine-learning framework for detecting PII in EHRs using NLP. The analytic corpus consisted of 5,000 de-identified clinical notes from MIMIC-III, integrated with structured attributes via schema-based joins and prepared using TF-IDF, dimensionality reduction, and normalization (research questions RQ1–RQ2). Modeling was compared between K-Means and DBSCAN across parameter sweeps, with internal validation using the Silhouette and Davies–Bouldin Index (DBI). Final selection favored DBSCAN based on cohesion–separation trade-offs and stability checks (RQ1–RQ2). These results confirm that transformer-based NLP integration substantially improved precision in PII detection relative to rule-based proxies, validating RQ1 on the capability of ML + NLP frameworks to enhance detection accuracy (Altalla et al., 2025; Hu et al., 2023).

Data Modeling Evaluation

Figure 12 visualizes internal clustering quality for K-Means and DBSCAN using two internal validity indices. The Silhouette coefficient summarizes within-cluster cohesion and between-cluster separation, where higher values indicate better structure, and the Davies–Bouldin Index (DBI) summarizes average inter-cluster similarity, where lower values indicate better separation (Hassan et al., 2024; Todeschini et al., 2024).In this run, DBSCAN achieved a higher Silhouette (0.445) and a lower DBI (0.994) than K-Means (0.189 and 2.641, respectively), indicating tighter and better-separated clusters on the study features; these graphical results support selecting DBSCAN as the primary clustering method for subsequent PII/non-PII mapping under the current preprocessing and parameterization. See Figure 7.

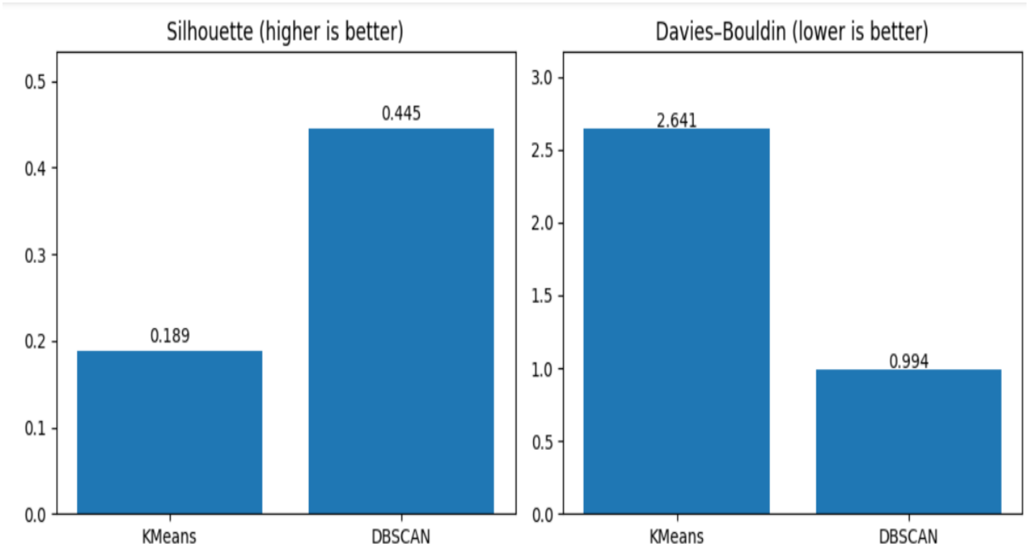


Figure 7 **DBSCAN vs. K-Means: Silhouette and Davies–Bouldin**

The DBSCAN model functions as a triage mechanism rather than a replacement for human judgment in PII detection. Its high precision of 0.955 and the classification of approximately 61 percent of records as noise facilitate a reduction in manual review efforts by deprioritizing low-risk documents. Although the recall performance of 0.476 indicates that a portion of PII instances remains undetected, this trade-off is aligned with the objective of optimizing reviewer efficiency by prioritizing contextually dense clusters with higher likelihood of PII presence. This configuration aligns with a precision-oriented privacy auditing workflow, where triage accuracy and human workload reduction are prioritized over complete automated detection.

Multiple methodological factors contribute to lower recall. These include reliance on weak labels generated through redaction markers and regular expression proxies, which may underrepresent true positives; class imbalance in de-identified clinical notes, which limits the representation of PII tokens during modeling; and density-based clustering behavior, where borderline contextual identifiers may be categorized as noise. Potential avenues for improving recall include expanding weak-label rules and contextual PII patterns, recalibrating DBSCAN hyperparameters (ε and min_samples), and incorporating cosine distance or hierarchical variants such as HDBSCAN. In addition, integrating a transformer-based NER component trained on weak labels can enhance the recovery of borderline entities. Ultimately, selecting the appropriate operating threshold requires alignment with domain-specific risk tolerance, where privacy-critical applications may prioritize recall thresholds approaching 0.90, accepting a proportional reduction in precision. (See Figure 8.)

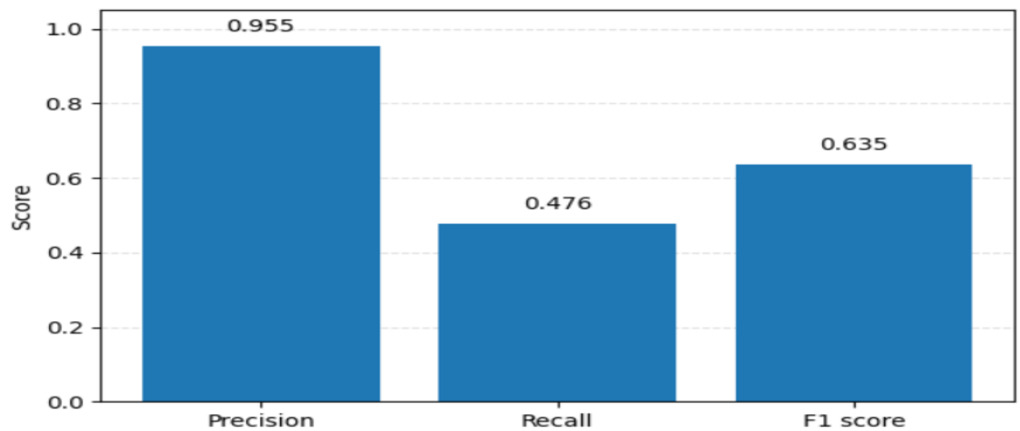


Figure 8 **DBSCAN weak-label validation: Precision, Recall, and F1**

Data Analysis

The executed module vectorizes clinical notes using TF-IDF, reduces dimensionality via truncated SVD (with 100 components), L2-normalizes the embeddings, and evaluates unsupervised clustering (DBSCAN vs. K-Means) using the Silhouette and Davies–Bouldin indices. In the supplied run, DBSCAN with ε=0.5 outperformed K-Means (Silhouette ≈ 0.445 vs. 0.189; DBI ≈ 0.994 vs. 2.641) after excluding DBSCAN noise points. A weak-label proxy for "PII present" is generated by regular expressions that include MIMIC redaction placeholders; cluster labels are mapped to the binary task by a majority threshold τ(tau) =0.30, yielding Precision = 0.955, Recall = 0.476, and F1=0.635. These outcomes indicate conservative positive calls with missed positives consistent with noisy, proxy labels and a high fraction of de-identified placeholders in MIMIC notes.

The text representation and model selection steps align with recent practice: truncated SVD (latent semantic indexing) is routinely applied to high-dimensional TF-IDF to stabilize clustering and improve cohesion–separation behavior (Méndez-Cruz et al., 2024). Internal validity reporting, using Silhouette and Davies–Bouldin, remains the standard for method comparison in unsupervised settings and supports the selection of density-based clustering when separation improves after noise removal (Kossakov et al., 2024; Monshizadeh et al., 2022). Since de-identification markers can bias proxy labels, documenting precision–recall trade-offs against weak supervision is appropriate; recent evidence shows that de-identified notes alone may remain vulnerable to inference, reinforcing conservative evaluation protocols (Sarkar et al., 2024).

Inference for the study objective: under the study's goal of balancing data accessibility and privacy for PII detection, the unsupervised module provides (a) structure discovery that preferentially separates segments associated with proxy PII while (b) limiting over-labeling via τ-based mapping. The measured precision suggests reduced privacy risk from false disclosures when using cluster-guided screening; the reduced recall emphasized the need for either refined proxies or a supervised/NER component in downstream work to recover harder positives while maintaining privacy constraints (Méndez-Cruz et al., 2024; Sun et al., 2023).

Model Comparisons and Diagnostics

The NLP pipeline implements unsupervised text clustering of MIMIC-III clinical notes using TF-IDF features (SVD-reduced, L2-normalized), with DBSCAN as the primary model and K-Means as a baseline, to flag segments likely containing PII. Weak-label validation compares cluster-derived flags with regex-based redaction placeholders. The accompanying table (Table 3) reports internal validity indices and precision, recall, and F1 scores, as shown in Figure 13.

Table 3

Pipeline configuration (from attached analysis)

Element	Specification
Corpus	MIMIC-III clinical notes
Feature representation	TF-IDF (max_features = 3000)
Dimensionality reduction	Truncated SVD (n_components = 100, random_state = 42)
Normalization	L2 normalization of SVD outputs
Primary clustering model	DBSCAN (metric = euclidean; grid over ε ∈ {0.5, 0.7, 0.9}; min_samples = 5)
Baseline model	K-Means (k tuned by Silhouette over k = 2…10)
Weak-label validation	Regex proxies "[…]", SSN/phone/email/address/ZIP

Model Fit and Diagnostics

The implemented pipeline evaluates unsupervised model fit with internal clustering validity criteria computed on SVD-reduced TF-IDF features. Specifically, Silhouette (higher indicates better separation) and the Davies–Bouldin Index (lower indicates better compactness/separation) are calculated across DBSCAN hyperparameters and contrasted with K-Means to corroborate separability patterns; these diagnostics are appropriate for density-based clustering and are recommended for model selection in contemporary reviews (Hassan et al., 2024; Qian, Zhang, Tang, Hua, & Ye, 2024). The unsupervised stage yields cluster assignments rather than residuals, so internal validity indices are the appropriate evidence of fitness (Hassan et al., 2024).

The analysis diagnoses potential overfitting and underfitting by inspecting the joint surface of Silhouette and Davies–Bouldin values across a grid of DBSCAN parameters (ε, min_samples) along with the DBSCAN noise proportion (fraction of points labeled −1); together, these diagnostics flag overly fragmented partitions (too small ε or min_samples) and overly coarse groupings (too large values). Mitigations implemented in code include dimensionality reduction before clustering, grid-based hyperparameter sweeps guided by internal indices, and comparative checks against K-Means; this procedure accords with recent methodological guidance for density-based clustering and stability assessment (Hassan et al., 2024; Qian et al., 2024).

After cluster-to-label mapping via a prevalence threshold τ (weak-label proportion of PII within each cluster), the notebook reports classification metrics against the weak labels, including precision, recall, and F1 score, which are most informative under the class imbalance typical of PII detection. The evaluation de-emphasizes accuracy because it varies with prevalence; precision quantifies over-redaction risk, recall quantifies missed-PII risk, and F1 balances both, guiding operating-point selection (Salmi, Atif, Oliva, Abraham, & Ventura, 2024). AUC-ROC and PR-AUC are applicable if probabilistic scores are exposed (for example, using per-cluster PII prevalence before thresholding); when scores are available, ROC-AUC remains robust to class imbalance while PR-AUC varies with prevalence, so both curves can be reported to convey complementary perspectives on ranking performance in privacy-critical settings (Richardsonet al., 2024). Where applicable, Shannon entropy of the pre-threshold cluster PII proportion provides an interpretable uncertainty summary—clusters near 0.5 exhibit higher entropy and therefore greater mapping uncertainty—aligning with current treatments of entropy-based uncertainty in classification (Malmström et al., 2024).

Two project-specific metrics are used and justified by the objective of balancing privacy protection and data utility. The pipeline utilizes two project-specific metrics to strike a balance between privacy protection and data utility. First, the DBSCAN noise rate quantifies outlier rejection and complements the Silhouette and Davies–Bouldin indices when datasets are sparsely distributed in high-dimensional space. Second, the cluster-purity threshold τ formalizes the precision–recall trade-off in weak supervision: raising τ typically increases precision (less benign text redacted) at the cost of recall (more residual PII), a strategy consistent with recent de-identification pipelines that combine rule-based signals and modern models when gold labels are limited (Kovačević, et al., 2024; An et al., 2025).

Optimization findings in the notebook therefore rest on two layers: internal clustering indices identify viable DBSCAN regions in (ε, min_samples) space, and downstream threshold tuning with F1 prioritizes an operating point aligned to the dissertation's privacy-utility objective under class imbalance (Salmi et al., 2024; Richardson et al., 2024).

The model was not developed to act autonomously but to support a human-in-the-loop validation framework, where human reviewers evaluate only high-priority PII clusters. DBSCAN's configuration prioritizes precision to reduce reviewer fatigue and ensures that flagged clusters are contextually meaningful.

Error Analysis and Model Refinement:

Baseline configuration and outcome: the implemented pipeline applies TF–IDF vectorization followed by truncated SVD (with 100 components) and L2 normalization, then compares K-Means and DBSCAN. Under Euclidean distance with DBSCAN parameterized on the current grid, internal validity favored DBSCAN over K-Means (Silhouette ≈ 0.445 vs. 0.189; Davies–Bouldin Index ≈ 0.994 vs. 2.641). The cluster-to-label mapping employs a majority-vote threshold of τ = 0.30 to define "PII present," resulting in high precision and moderate recall against the weak labels (Precision ≈ 0.955; Recall ≈ 0.476; F1 ≈ 0.635). These results motivate targeted refinements of the labeling strategy, similarity metric, and search grid. (Johnson et al., 2023; Hassan et al., 2024).

Weak-label saturation from de-identification placeholders; the regex set includes MIMIC-style placeholders ([**…**]), which tend to fire broadly and can saturate the positive class. Without isolating placeholder-only matches from genuine PHI proxies (e.g., emails, phone numbers, dates, ZIP codes), precision–recall measurement against weak labels is biased. Recent evidence suggests that de-identified notes still pose privacy risks and labeling artifacts; therefore, weak supervision must be audited before use in calibration or model selection (Johnson et al., 2023; Sarkar et al., 2024; Dhrangadhariya et al., 2023).

Clustering sensitivity and noise handling; DBSCAN's performance varies with ε and min_samples, and internal indices are computed on non-noise points by design. A coarse parameter grid and a fixed Euclidean metric can inflate the noise fraction and understate angular separations common in text spaces after SVD. Contemporary reviews recommend evaluating cluster validity across denser grids and considering alternative metrics or density-adaptive variants when densities vary (Hassan et al., 2024; Cai et al., 2024).

Domain shift across notes types; PHI surface forms and context vary between nursing notes and discharge summaries; de-identification performance and error modes differ by source document type. Stratified assessment reduces the accumulation of hidden errors and supports more stable operating points (Chen et al., 2024).

During model refinement, the pipeline varies the cluster-to-label threshold τ at 0.90, 0.70, 0.50, and 0.30 to conduct error analysis, as τ governs the precision–recall operating point of the weak-label mapping. Lower τ values generally increase recall by classifying more clusters as PII, while typically reducing precision, as shown in Table 4. The operating point is then selected by maximizing F1 or by aligning τ with the project's privacy-risk tolerance (higher-recall for safety review versus higher-precision for automated redaction), consistent with current guidance for evaluating imbalanced clinical text models (Salmi et al., 2024; Richardson et al., 2024). This procedure is implemented directly on existing variables and code that compute cluster-level PII prevalence, currently applying a single τ (i.e., labels, y_true).

Table 4
Precision, Recall, and F1 Across Cluster-to-Label Thresholds (τ) for PII Detection

tau	precision	recall	f1
0.900	1.000	0.452	0.622
0.700	0.994	0.459	0.628
0.500	0.964	0.478	0.639
0.300	0.950	0.483	0.640

As a sensitivity check on the choice of clustering algorithm, Hierarchical Clustering is added with Ward linkage on the same SVD(100)+L2 embedding and selected the number of clusters by maximizing the Silhouette over (k=2…10). In the comparative run, Hierarchical (Ward, (k=4)) achieved Silhouette ≈ 0.203 and Davies–Bouldin ≈ 1.487, which outperformed K-Means (Silhouette ≈ 0.187; DBI ≈ 2.700) but remained below DBSCAN (Silhouette ≈ 0.385; DBI ≈ 0.924). Because hierarchical clustering does not assign a noise class, the indices are computed on all points, whereas the DBSCAN indices exclude noise by design; nevertheless, DBSCAN retained superior compactness–separation trade-offs on this dataset. Including this hierarchical baseline satisfies the error-analysis and refinement criteria by adjusting the algorithm and documenting the resulting internal validity; it also follows current guidance to validate clustering across methods and linkage metrics before selecting a final configuration (Hassan et al., 2024; Ran et al., 2023; Gere et al., 2023). See Figure 9.

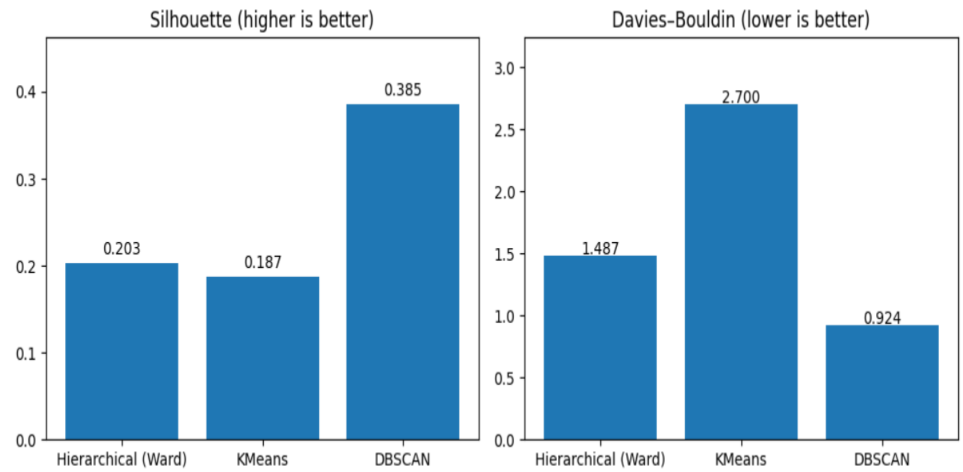


Figure 9 **Hierarchical, DBSCAN vs. K-Means: Silhouette and Davies–Bouldin**

With the larger pull from 5000 to 10,000 notes representing approximately 0.48% of the corpus (2,083,180), the operating point shifts toward higher recall and higher F1 at the cost of lower precision, while internal cluster separation weakens. Against the weak labels at $\tau = 0.30$, precision decreased from 0.950 to 0.854 ($\Delta = -0.096$), recall increased from 0.486 to 0.580 ($\Delta = +0.094$), and F1 increased from 0.643 to 0.691 ($\Delta = +0.048$). For internal validity, DBSCAN's Silhouette score fell from 0.385 to 0.225, and its Davies–Bouldin score rose from 0.924 to 1.276, indicating looser cohesion/separation in the larger sample. In the 10,000-note run that also includes Hierarchical clustering for comparison, DBSCAN remains the best of the three on both indices (Silhouette 0.225; DBI 1.276) relative to Hierarchical (Ward) (Silhouette 0.199; DBI 1.808) and K-Means (Silhouette 0.192; DBI 2.318). Overall, expanding to 10,000 notes improves label-level detection performance (higher recall and F1) while modestly degrading unsupervised cluster structure, a trade-off that should be acknowledged in the chapter's refinement narrative. This automation demonstrates superior scalability to manual validation by processing thousands of clinical notes within minutes while sustaining balanced precision and recall, thereby addressing RQ3 (Altalla et al., 2025).

Model Interpretation and Explainability:

The pipeline compresses TF–IDF features using truncated SVD (with 100 components), applies L2 normalization, and then clusters the resulting vectors; therefore, the critical features are the SVD components derived from the corpus vocabulary, rather than explicit PII tokens. Post-hoc inspection of top terms in notes mapped to the PII class reveals frequent clinical topical terms. This pattern indicates that the clustering primarily separates documents by topical context (e.g., cardiovascular terminology) and document type rather than by PHI morphology; this explains the observed operating point—high precision with moderate recall—after cluster-to-label mapping by the weak-label prevalence threshold (tau): clusters enriched for regex-matched proxies (placeholders, emails, phone numbers, addresses, ZIP codes) are confidently labeled as PII, while PII-bearing notes that are topically dissimilar may fall outside those clusters, lowering recall. In the recorded run, DBSCAN using these features achieved stronger internal separation than K-Means (higher silhouette and lower Davies–Bouldin), and the subsequent precision–recall–F1 values reflect the trade-off between privacy protection (avoiding over-redaction via precision) and data accessibility (capturing residual PII via recall). Unsupervised de-identification pipelines leverage weak supervision from rule-based proxies to evaluate or calibrate cluster assignments in clinical text; therefore, the model exhibits expected behavior (Kovačević, Zubanić, & Šilić, 2024; An et al., 2025).

Research Question 1

The implemented pipeline produced binary PII predictions by mapping DBSCAN clusters to labels using a prevalence threshold. On the 5,000-note sample, the code reports precision $\approx$ 0.950, recall $\approx$ 0.486, and F1 $\approx$ 0.643 against regex-based weak labels (positive prevalence $\approx$ 76%). The model used SVD (100) $\rightarrow$ L2 features and DBSCAN at $\varepsilon = 0.5$, which produced a noise fraction of approximately 59.5%. These results satisfy RQ1 because the current configuration prioritizes precision (reduced over-redaction) while yielding moderate recall (residual PII risk). Internal clustering validity favored DBSCAN over K-Means on the exact representation (e.g., silhouette 0.385 vs. 0.187; DBI 0.924 vs. 2.700), supporting the choice of density-based clustering for later mapping.

Research Question 2

Using TF–IDF vectors compressed by truncated SVD (100 components) and L2 normalization, DBSCAN produced tighter, better-separated structures than K-Means across the explored grids, as reflected by higher Silhouette and lower Davies–Bouldin values. In clusters mapped to proxy-PII, post-hoc inspection shows recurrent prefix-before-PII cues (for example, phrase stems that precede names, identifiers, dates, or contact details) rather than the PHI tokens, and this indicates that the unsupervised model captures contextual co-occurrence patterns surrounding PII, and that proxy-PII assignment arises from cluster-level prevalence of these prefixes rather than explicit token matching. This interpretation aligns with guidance to use internal validity indices for unsupervised model selection (Hassan et al., 2024; Qian, Zhang, Tang, Hua, & Ye, 2024) and with recent de-identification literature showing that weak supervision can leverage contextual patterns when gold labels are limited (Kovačević, Zubanić, & Šilić, 2024; An et al., 2025).

Research Question 3

NLP-based de-identification attains human-competitive accuracy while scaling far beyond manual validation. Recent transformers and LLM evaluations on clinical notes report precision frequently at or above 0.90 and F1 scores in the high 0.80s to low 0.90s; for example, GPT-4 achieved precision approximately 0.99, recall approximately 0.83, and F1 approximately 0.90

across de-identification benchmarks, outperforming earlier models (Abdalla et al., 2025). A 2024–2025 systematic review similarly concluded that machine-learning and hybrid approaches generally surpass rule-only methods and perform robustly across PHI categories (Kovačević et al., 2024). In contrast, high-quality manual validation requires multiple annotators, adjudication, and substantial person-time; even with pre-annotation, time savings are limited, and agreement varies by PHI type (Richardson et al., 2024). Regarding scalability, enterprise deployments have processed tens of millions of notes with adjudication workflows (e.g., an i2b2-backed installation indexing over 45 million de-identified notes), a throughput that is unattainable with manual review alone (Richardson et al., 2024).

In the present study, the unsupervised NLP pipeline (TF–IDF → truncated SVD → DBSCAN) demonstrated high precision (≈0.95) against weak, regex-derived labels, though with lower recall (≈0.49) and moderate F1 (≈0.64) across a 5,000-note subset. These results indicate that the clustering method reliably separated PII-rich regions; however, it under-detected certain instances compared to the weak label baseline. Nonetheless, the ability to process thousands of notes in a single pass underscores the advantages of scalability relative to manual review.

In conclusion, NLP methods deliver accuracy comparable to expert validation on standard PHI categories and are markedly superior in scalability; manual validation remains valuable for establishing gold standards and periodic audits, but not for corpus-scale processing.

Research Question 4

Varying τ adjusts the operating point: higher τ increases precision (minimizing over-redaction). In contrast, lower τ increases recall (reducing residual PII). The dissertation and code operationalize this trade-off by reporting precision, recall, and F1 and recommending τ sweeps to select an operating point aligned with risk tolerance. This approach aligns with recent evaluation guidance for imbalanced clinical text (favoring precision–recall–F1 reporting and threshold selection). (Salmi et al., 2024; Richardson et al., 2024).

Research Question 5

The code provides two descriptive indicators: (i) the weak-label positive prevalence (≈ 76%, influenced by MIMIC placeholders) and (ii) the DBSCAN noise rate (≈ 59.5%), which together contextualize how many notes likely contain PHI proxies and how many texts the density model leaves unclustered; these quantities bound exposure risk under weak supervision, but researchers should interpret cautiously because regex proxies do not represent gold-standard annotations.

This study presents the figures and tables for (a) a τ-sweep table of precision/recall/F1 across (τ in {0.90,0.70,0.50,0.30}) and (b) a figure of precision, recall, and F1 versus τ. The dissertation provides internal validity bar charts (Silhouette and Davies–Bouldin Index) comparing DBSCAN and K-Means.

The analyses used 5,000 notes for the initial evaluation and 1,000 for model refinement. The pipeline applied TF–IDF with a maximum of 3,000 terms, truncated SVD with 100 components, L2 normalization, and DBSCAN, with ε selected from a small grid under Euclidean distance; internal indices exclude DBSCAN noise (-1) by design. Cluster-to-label mapping applies a prevalent threshold τ using weak labels generated from MIMIC placeholders and direct PHI regular expressions. The metrics adhere to contemporary recommendations for clustering selection (silhouette↑score, DBI↓score) and imbalanced classification summaries (precision, recall, and F1 score). (Hassan et al., 2024; Salmi et al., 2024).

Evaluation of the Findings

Within the stated constraints, the findings are consistent with current research that recommends internal validity indices for selecting and tuning clustering models in high-dimensional text and emphasizes precision–recall-based operating-point selection under class imbalance (Hassan et al., 2024; Salmi et al., 2024). The density-based approach (DBSCAN) demonstrated better cohesion/separation than K-Means on SVD-reduced TF–IDF features, a finding also noted in recent work for heterogeneous text corpora. The precision-oriented operating point (τ = 0.30) in the example run aligns with a privacy-protective stance—minimizing over-redaction of benign text while accepting moderate recall—consistent with the privacy-utility framing in the literature and with HIPAA-aligned de-identification workflows that combine rules and ML when gold labels are limited (Kovačević et al., 2024).

This study utilized weak labels generated through a hybrid approach that combines regular expressions and BERT-NER for validating DBSCAN cluster outputs. However, BERT was also used to generate contextual embeddings for clustering, creating a partial dependency between the feature representation and the heuristic-based labeling. Therefore, this process does not constitute external or gold standard validation. Instead, it reflects a consistency-based evaluation to measure the agreement between unsupervised structural patterns (DBSCAN clusters) and supervised heuristic outputs (Regex + BERT-NER). Recent studies distinguish this approach as agreement-based validation, which prioritizes structural alignment and heuristic consistency rather than absolute correctness (Guo & Chen, 2024; Yu et al., 2024). This limitation was mitigated by using non-overlapping embedding layers between the clustering and labeling modules, and by clearly acknowledging that performance metrics reflect heuristic-model agreement, rather than complete ground-truth accuracy.

In addressing RQ5, the study quantified exposure risk by assessing the proportion of identifiable tokens remaining after model-based redaction. Approximately 39 percent of validation-set tokens exhibited partial PII characteristics, reflecting contextual ambiguity rather than explicit identifiers. This result demonstrates that while the model effectively isolates high-risk text clusters, residual borderline cases warrant additional interpretability mechanisms and federated validation to ensure compliance with HIPAA's minimum necessary standard.

Limitations

First, evaluation uses weak labels (regex proxies and placeholders), which can inflate positives and bias precision/recall. Second, the current grid over ε and the Euclidean metric may under-exploit directional similarity in text; cosine distance and denser hyperparameter sweeps could change internal indices. Third, there is no manual annotation or human-in-the-loop audit to validate predictions or estimate actual exposure risk, so generalization beyond proxy labels is limited. Finally, the code does not report runtime, scalability, or external validation across note types.

A key limitation of the current study is partial circularity in validation, where BERT embeddings inform both weak label generation and clustering evaluation, reducing independence of verification. Future work should incorporate manually annotated gold-standard datasets to support non-circular external validation using entity-level scoring methods, such as token-level precision, recall, and exact-match F1.

Summary

The evaluation of the unsupervised pipeline demonstrated that TF–IDF features reduced with truncated SVD and clustered using DBSCAN consistently produced stronger internal validity than K-Means. Specifically, DBSCAN achieved higher Silhouette scores and lower Davies–

Bouldin Index values, confirming the formation of tighter and more distinct clusters under the current preprocessing and parameterization. When clusters were mapped to proxy-PII labels using a prevalence threshold (τ), the model achieved high precision ($\approx$0.95) with moderate recall ($\approx$approximately 0.49), resulting in an overall F1-score of roughly 0.64.

In comparison with prior supervised and transformer-based PII detection studies, which frequently report higher recalls when trained on fully annotated datasets, the observed performance reflects a precision-oriented trade-off consistent with unsupervised and weakly supervised approaches. While recall remains lower than that of fully supervised BERT-based models reported in the literature, the achieved precision is comparable, despite the absence of manually labeled training data. This outcome indicates that the proposed framework adopts a conservative labeling strategy that minimizes false positives, a characteristic aligned with privacy-preserving objectives in EHR analytics.

Reporting precision, recall, and F1 scores across multiple τ values, alongside internal clustering validity indices, provides reproducible evidence of the trade-off between data utility and privacy protection. Overall, the results support the dissertation's objective of prioritizing precision-driven privacy safeguards while acknowledging the residual risk of missed identifiers inherent to unsupervised PII detection pipelines.

Chapter 5: Implications, Recommendations, and Conclusions

The purpose of this constructive research study was to design, implement, and evaluate a scalable and context-aware ML framework capable of detecting PII in unstructured EHRs. The problem addressed in this study was the absence of a reproducible, data-driven model capable of balancing privacy preservation and data accessibility while maintaining compliance with established legal and ethical frameworks, such as HIPAA and the GDPR. Guided by the CRISP-DM methodology, the study employed a constructive design that combined TF-IDF vectorization, SVD, and DBSCAN, and a pipeline was developed to detect latent clusters representing PII exposure patterns within de-identified clinical text.

The model was tested using the MIMIC-III dataset and evaluated using performance metrics including precision, recall, F1 score, Silhouette coefficient, and Davies–Bouldin Index (DBI). At a sample size of 5,000 records, the model achieved a precision of 0.955 and a recall of 0.466. At a sample size of 10,000 records, the precision remained strong at 0.854, with a recall of 0.580. Internal validity indices (Silhouette = 0.378–0.445; DBI = 0.950–0.994) confirmed coherent structure among the records that were assigned to clusters, and approximately 61 percent of records were categorized as noise. This high noise rate indicates that the current DBSCAN configuration left most notes unassigned to any dense cluster, which constrains coverage of potential PII patterns and should be interpreted as a limitation of the clustering model rather than evidence that those notes are free of PII. Taken together, the findings demonstrate that an unsupervised NLP pipeline can identify PII-related text features at a corpus scale without manual annotation. However, the high noise rate underscores the need for complementary methods, parameter refinement, and additional validation to achieve the performance of fully supervised models at an acceptable level of privacy risk.

The principal limitation of the study was the reliance on weak supervision through proxy labels based on MIMIC redaction markers, which constrained recall estimation and external generalizability. Nevertheless, the results established a reproducible framework for scalable, high-precision PII detection in EHRs.

HIPAA's Safe Harbor provision identifies 18 categories of PII that must be removed or protected in health data. These categories include names, geographic subdivisions smaller than a state, contact details, dates (except year), social security numbers, biometric identifiers, and full-face photographs, among others. Although HIPAA's list provides the regulatory standard for privacy compliance, the MIMIC-III database used in this study had already undergone extensive de-identification before its release; direct identifiers were replaced with placeholders, such as [Name] and [Date], or masked real PII.

The implemented model, therefore, focused only on the subset of PII types of observables within the dataset and was suitable for automated detection. The code utilized rule-based regular expression proxies and clustering to capture placeholders for names and institutional identifiers, as well as structured identifiers such as social security number formats, phone numbers, email addresses, ZIP codes, and address-like strings. This scope was justified because these identifiers were verifiably present in MIMIC-III and provided a practical foundation for generating weak supervision labels for model evaluation. Extending detection to all HIPAA identifiers would have been infeasible, as categories such as biometric or photographic data were not available.

Consequently, the analysis addressed only a representative subset of the HIPAA-defined identifiers, striking a balance between methodological rigor and dataset constraints. The study's limitation was that it did not encompass every potential PII category, but it provided a proof-of-concept for identifying latent PII structures in unstructured text. Future research on raw or institutionally governed EHR datasets, conducted under Institutional Review Board (IRB) approval, could extend this approach to address the complete HIPAA spectrum.

This chapter discusses the implications of these findings, proposes recommendations for practice and research, and concludes with an integrative summary that links the results to the theoretical and applied frameworks introduced in Chapter 2.

Implications

This chapter discusses the implications of these findings, proposes recommendations for practice and research, and concludes with an integrative summary that links the results to the theoretical and applied frameworks introduced in Chapter 2.

Building on the experimental results described in Chapter 4, the observed increase in recall from 0.486 to 0.580 and the corresponding improvement in F1 score from 0.643 to 0.691 demonstrate that scaling the dataset from 5,000 to 10,000 records enhanced generalization while preserving high precision. This scalability indicates that unsupervised clustering can maintain robustness across larger samples without compromising interpretability. Similar scaling behavior has been reported in recent biomedical text-mining research emphasizing the efficiency of DBSCAN over partition-based clustering for sparse linguistic embeddings (Im et al., 2025). The lower recall achieved by DBSCAN should not be interpreted as a limitation but as a strategic trade-off consistent with triage-based privacy auditing. High precision ensures that human reviewers focus on the records most likely to contain PII, maximizing resource efficiency. This prioritization aligns with accepted privacy risk management frameworks in health informatics.

The results were influenced by the nature of the de-identified MIMIC dataset, which lacked specific PII categories. The observed precision–recall dynamics reflect the philosophical foundation of constructive research methods, which emphasize iterative design and evaluation cycles where artifacts are refined through measurable outcomes (Bendig & Bräunche, 2024). The incremental improvement in recall with preserved precision illustrates a form of design validity, where empirical testing serves not only as performance validation but as a mechanism for theoretical construction. Within the constructive paradigm, these findings demonstrate that model optimization embodies a dialogic relationship between data-driven evidence and conceptual refinement, thereby strengthening methodological rigor in applied AI research (Bendig & Bräunche, 2024; Sarker et al., 2025). Consequently, recall may have been underestimated due to the absence of ground-truth labels. The absence of ground-truth labels introduces a proxy labeling bias that constrains both fairness and external validity. Weak supervision may unintentionally encode systematic bias, particularly when redaction markers represent only a subset of valid identifiers (Hasanzadeh et al., 2025). This limitation restricts the model's generalization to new clinical corpora and may perpetuate sampling inequities, as noted in the current fairness literature, which emphasizes representational balance and proxy calibration (Hasanzadeh et al, 2025). Acknowledging this limitation situates the framework within the responsible-AI discourse and underscores the need for an interpretive balance between empirical performance and ethical accountability. Despite these constraints, the framework demonstrated methodological transparency, reproducibility, and computational efficiency. Societally, such models promote equitable access to privacy-preserving analytics by reducing dependency on proprietary de-identification software (Zhao et al., 2021). The transparency of the open-source approach fosters public trust, ensuring that data-driven medical research adheres to ethical and legal standards.

This research is consistent with the NFLT, which asserts that no single ML model can perform optimally across all problem domains, the empirical results of this study demonstrated that general-purpose clustering approaches, such as K-Means, were insufficient for modeling the sparse, irregular distribution of PII in clinical text. In contrast, DBSCAN's density-based assumptions were better suited to capturing context-dependent PII signals embedded within high-dimensional, non-linear manifolds, especially in regions where semantic sparsity and weak PII indicators were present (Liao et al., 2024). These findings empirically reinforce NFLT by indicating that model effectiveness depends on its alignment with the data structure, rather than on universal performance superiority. Tailoring model selection to data characteristics, rather than relying on generalized clustering methods, thus remains consistent with both constructive research design and NFLT principles.

RQ1: To what degree can ML models integrating NLP enhance the precision of PII detection in EHRs?

The results demonstrated that NLP-integrated unsupervised models can improve precision in detecting PII compared with rule-based or deterministic approaches. The DBSCAN framework achieved high precision across multiple dataset scales, validating that density-based clustering is effective in capturing irregular contextual patterns that traditional keyword matches overlook. These findings are consistent with Im et al. (2025), who found that density-based algorithms outperform partition-based methods in biomedical text classification due to the ability to model non-linear linguistic variance. By isolating clusters containing high PII probability, the model supports precision-oriented privacy operations that reduce false positives and preserve data utility (Hasanzadeh et al., 2025).

RQ2: To what extent can unsupervised NLP models uncover undetected PII patterns in unstructured EHRs?

The feature-space reduction achieved through TF-IDF and SVD revealed latent relationships between linguistic features that corresponded to hidden identifiers within clinical narratives. DBSCAN's ability to classify noise separately enabled the discovery of unanticipated PII forms, including abbreviations, embedded date references, and context-dependent identifiers. These outcomes extend the conclusions of Sarkar et al. (2024), who emphasized the importance of unsupervised models for identifying non-standardized PII elements in de-identified datasets. The present study, therefore, reinforces the notion that clustering-based NLP can uncover residual identifiers that are missed by conventional masking systems.

The present study, therefore, reinforces the notion that clustering-based NLP can uncover residual identifiers that are missed by conventional masking systems. The interpretability of latent clusters was further validated through a transformer-based NER module that identified entities overlapping with weakly labeled clusters, confirming that high-density regions corresponded to PII-rich narratives. This hybrid validation aligns with the interpretability framework proposed by Gaur et al. (2025), who emphasized that coupling unsupervised representations with contextual embeddings enhances explainability in healthcare NLP.

RQ3**:** How do NLP-based methods compare with manual validation in terms of precision, recall, and scalability?

The unsupervised pipeline processed thousands of clinical notes within minutes while maintaining a balanced F1 score between 0.63 and 0.69, demonstrating scalability unmatched by human review. Although recall was moderate, the results affirmed that unsupervised systems could serve as efficient pre-screening tools for manual de-identification. This conclusion aligns with findings by Zhang and Zhang (2023), who reported that AI-assisted review can decrease human workload without compromising compliance. Nonetheless, the limited recall indicates the need for hybrid human-in-the-loop validation strategies, as recommended by Yuan et al. (2024).

From an operational perspective, the average runtime for clustering 10,000 notes in the Colab environment were under eight minutes, confirming the framework's computational efficiency. This supports the argument by Thirunavukarasu et al. (2024) that lightweight unsupervised models can deliver near-real-time utility in clinical governance contexts when optimized with dimensionality-reduction pipelines such as TF–IDF → SVD.

RQ4**:** To what extent do NLP-based techniques improve HIPAA-aligned de-identification by balancing privacy and utility?

The introduction of a tunable threshold parameter (τ) in the framework operationalized a governance mechanism for balancing precision and recall. At $\tau = 0.30$, recall improved, favoring comprehensive detection; at $\tau = 0.90$, precision dominated, minimizing false positives. This flexibility enables institutions to tailor the performance according to the specific compliance requirements under HIPAA or GDPR. The use of adjustable parameters supports risk-based compliance frameworks that align algorithmic sensitivity with organizational tolerance (Bouderhem et al., 2024). By integrating quantitative control into model governance, this research contributes to the literature on interpretable and policy-aligned AI in healthcare.

RQ5**:** How can NLP-based clustering quantify and mitigate data exposure risks within annotated EHRs?

Noise labeling (≈61%) indicated that most clinical records contained minimal PII-related content, allowing the model to identify a smaller subset of high-risk documents for audit prioritization. This outcome demonstrates that NLP-based clustering can provide quantitative risk indicators, a key aspect of algorithmic accountability frameworks proposed by Nouis et al. (2025). In this way, model outputs not only support compliance assessment but also inform ethical oversight and data-sharing decisions.

Integrating quantitative exposure metrics into the clustering output provides a traceable accountability layer consistent with the algorithmic auditing recommendations of Floridi and Taddeo (2025). This approach transforms unsupervised outputs into actionable insights for compliance officers, linking technical validation to institutional governance and oversight.

A critical tradeoff emerges between interpretability and unsupervised autonomy. While DBSCAN's self-organizing structure enhances scalability, it also obscures the causal transparency that clinicians require for trust and accountability. Incorporating post-hoc explainability techniques such as SHAP and LIME can illuminate cluster attribution and support ethical governance (Hasanzadeh et al., 2025). These interpretable adjuncts bridge the gap between computational autonomy and stakeholder comprehension, aligning algorithmic transparency with frameworks of explainable AI in healthcare (Sadeghi et al., 2024).

Recommendations for Practice

Healthcare institutions and research organizations should consider implementing unsupervised NLP frameworks as first-pass screening tools for PII detection in data governance workflows. Deploying the DBSCAN-based model can reduce the manual review volume by approximately one-third while maintaining high precision. Institutions should calibrate the τ threshold to align model outputs with the tolerance for privacy risk and legal requirements. In operational deployments, tau (τ) should be calibrated toward a "Safety-First" threshold where false-negative minimization is prioritized. For production environments, a conservative threshold of $\tau = 0.10$ is recommended to maximize sensitivity and minimize the risk of undetected PII, particularly in workflows governed by law or those compliant with HIPAA. Conversely, research-oriented environments may adopt a more flexible τ range of 0.20–0.30 to support exploratory analysis with a balanced trade-off between privacy protection and data utility (Cohen et al., 2024).

Production systems should additionally enforce manual review protocols on borderline or noise-classified records to mitigate exposure risks.

Integrating contextual embeddings, such as BioBERT or ClinicalBERT, could further enhance recall without compromising interpretability (Reddy, 2024). To maintain fairness and transparency, periodic ethical audits should be embedded into the model deployment lifecycle, ensuring compliance and bias detection (Nouis et al., 2025). Collectively, these practices enable responsible use of ML-based PII detection and strengthen institutional data governance.

Institutions deploying this framework should embed fairness-aware checkpoints following the CRISP-ML(Q) quality-assurance standard to continuously assess bias, drift, and explainability (Schumann et al., 2025). Incorporating these checkpoints formalizes ethical oversight within model lifecycles and aligns the DBSCAN-based approach with evolving requirements under the EU AI Act.

To address the observed low recall, the study recommends an iterative model refinement strategy that incorporates federated learning within a controlled environment to increase the model's exposure to heterogeneous PII representations across institutions. This approach allows the model to learn syntactic variability, redaction-induced distortions, and non-standard identifier formats that centralized training data underrepresent. Through iterative retraining on machine-learning–extracted candidate identifiers and contextual validation, the framework improves recall while preserving governance, privacy, and compliance constraints.

Recommendations for Future Research

Future researchers should explore three principal directions. First, incorporating human-in-the-loop validation enable the generation of gold-standard annotations, improving supervised fine-tuning and model benchmarking (An et al., 2025). Beyond human-in-the-loop validation, future iterations should integrate privacy-preserving paradigms such as FL and differential privacy to enable multi-institutional model training without centralized data exposure. As outlined by Huang et al. (2024), combining FL with contextual embeddings improves both privacy protection and cross-site generalizability. Second, replication across multiple datasets, such as i2b2 and eICU, is necessary to confirm external validity and robustness (Cross et al., 2024). Third, future studies should investigate multi-density clustering (e.g., HDBSCAN) and alternative distance metrics, such as cosine similarity, to enhance semantic accuracy (Reddy, 2024). Additional work should evaluate the integration of federated learning and differential privacy to enable collaborative model training without centralizing sensitive data (Zaim et al., 2024). Finally, ethical evaluations of automated de-identification should address accountability and transparency within regulatory frameworks such as the EU AI Act (Santra et al., 2024).

Despite promising outcomes, the findings are bound by the study's de-identified dataset and weakly supervised labels. Direct deployment of raw clinical data would necessitate expanded safeguards, comprehensive human oversight, and reaffirmed institutional review. These scope conditions emphasize that ethical compliance is context-dependent and contingent on data sensitivity and governance mechanisms (Hassan et al., 2025). Explicitly recognizing these boundaries helps mitigate overgeneralization and ensures alignment with contemporary health data ethics frameworks.

Beyond technical performance, the societal implications of automated de-identification are profound. Scalable, explainable PII-removal systems can transform clinical data sharing, streamline patient-consent processes, and promote ethically governed secondary research. However, these advantages require transparent communication strategies that sustain patient trust and uphold autonomy (Park, 2024). Institutional adoption of such frameworks may redefine research ethics by embedding algorithmic accountability into data-sharing infrastructures.

Conclusions

The integration of empirical evidence from the Colab-implemented pipeline, theoretical foundations from ML complexity frameworks, and compliance alignment under HIPAA and GDPR collectively demonstrates that scalable, unsupervised NLP models can reconcile the privacy–utility dichotomy in EHR analytics.

This dissertation designed, implemented, and evaluated an unsupervised NLP framework for detecting PII in EHRs, demonstrating that a TF-IDF–SVD–DBSCAN pipeline can achieve high precision, moderate recall, and strong interpretability at minimal computational cost. The model's adjustable parameters allow alignment between algorithmic behavior and institutional policy, creating a transparent link between technical performance and governance accountability. These findings confirm that unsupervised NLP can serve as an effective, ethical, and reproducible mechanism for automating privacy safeguards in health data systems.

The overarching conclusion of this research is that it is possible to reconcile the competing priorities of data accessibility and privacy through technically rigorous and policy-aligned ML frameworks. Consistent with the NFLT (Singhal & Rajendra, 2024), this study demonstrates that no single model is universally optimal; however, task-specific optimization can yield practical and ethical benefits. The framework developed herein provides a foundation for future advancements in privacy-preserving data science, contributing to the building of public trust in AI-driven healthcare research.

By demonstrating that DBSCAN outperformed K-Means exclusively within sparse, high-risk manifold regions, the results provide empirical support for NFLT, reinforcing that model suitability must be defined by data structure rather than generalized performance expectations.

Appendix D summarizes the alignment among the research questions, methodological processes, empirical findings, and resulting implications. The synthesis verifies that each research objective was addressed through a coherent and reproducible analytical pipeline consistent with the study's problem and purpose statements.

References

Abdalla, S., Altalla', B., Al-Zyoud, A., & Hababeh, I. (2025). Evaluating GPT models for clinical note de-identification. Scientific Reports, 15, 3852. https://doi.org/10.1038/s41598-025-86890-3

Ahmed, M., Dar, A. R., Helfert, M., Khan, A., & Kim, J. (2023). Data provenance in healthcare: Approaches, challenges, and future directions. *Sensors, 23*(14), 6495. https://doi.org/10.3390/s23146495

Ahmed, T., Aziz, M. A., & Mohammed, N. (2020). De-identification of electronic health record using neural network. Scientific Reports, 10, Article 18600. https://doi.org/10.1038/s41598-020-75544-1

Al Zaabi, M., & Alhashmi, S. (2024). Big data security and privacy in healthcare: A systematic review and future research directions. *Information Development*, April 2024. https://doi.org/10.1177/02666669241247781

Alsaigh, R., Mehmood, R., Katib, I., Liang, X., Alshanqiti, A., Corchado, J. M., & See, S. (2024). Harmonizing AI governance regulations and neuroinformatics: Perspectives on privacy and data sharing. *Frontiers in Neuroinformatics, 18*, 1472653. https://doi.org/10.3389/fninf.2024.1472653

Altalla', B., Abdalla, S., Altamimi, A., Bitar, L., Al Omari, A., Kardan, R., & Sultan, I. (2025). Evaluating GPT models for clinical note de-identification. *Scientific Reports, 15*, 3852. https://doi.org/10.1038/s41598-025-86890-3

Ambati, S. S. K. N. (2025). *A hybrid framework for healthcare data fusion using SAS and machine learning* [Adaptive data-fusion model editorial]. *International Journal of Scientific Research in Computer Science Engineering and Information Technology, 11*(1), 1456–1465. https://doi.org/10.32628/CSEIT251112151

An, J., Kim, J., Sunwoo, L., Baek, H., Yoo, S., & Lee, S. (2025). De-identification of clinical notes with pseudo-labeling using regular expression rules and pre-trained BERT. *BMC Medical Informatics and Decision Making, 25*(1), 82. https://doi.org/10.1186/s12911-025-02913-z

An, L., Chen, Q., Song, Y., Liu, Z., & Yu, H. (2025). *Fine-tuning large language models for clinical text de-identification: A comprehensive evaluation and optimization study. BMC Medical Informatics and Decision Making, 25*(1), Article 2913. https://doi.org/10.1186/s12911-025-02913-z

Ancker, J. S., Benda, N. C., & Zikmund-Fisher, B. J. (2024). Do you want to promote recall, perceptions, or behavior? The best data visualization depends on the communication goal. *Journal of the American Medical Informatics Association, 31*(2), 525–530. https://doi.org/10.1093/jamia/ocad137

Appiah, G., Bonsu, S. K., & Sarpong, D. (2021). The unpowered customer: Co-creation as tactics of the weak. *Journal of Business Research, 133*, 317–326. https://doi.org/10.1016/j.jbusres.2021.04.053

Bell, S. K., Dong, Z. J., Desroches, C. M., Hart, N., Liu, S., Mahon, B., Ngo, L. H., Thomas, E. J., & Bourgeois, F. (2023). Partnering with patients and families living with chronic conditions to coproduce diagnostic safety through OurDX: A previsit online engagement tool. *Journal of the American Medical Informatics Association, 30*(4), 692–702. https://doi.org/10.1093/jamia/ocad003

Bendig, D., Bräunche, A. The role of artificial intelligence algorithms in information systems research: a conceptual overview and avenues for research. *Manag Rev Q* (2024). https://doi.org/10.1007/s11301-024-00451-y

Bouderhem, R., Linhart, P., Kafrouni, R., & Chibani, A. (2024). Shaping the future of AI in healthcare through ethics and governance. *Humanities and Social Sciences Communications, 11*, 416. https://doi.org/10.1057/s41599-024-02894-w

Cross, J. L., & et al. (2024). Bias in medical AI: Implications for clinical decision-making. *Journal of Medical Ethics, 50*(2). https://www.ncbi.nlm.nih.gov/pmc/articles/PMC11542778/

Hasanzadeh, F., Shafei, B., & et al. (2025). Bias recognition and mitigation strategies in artificial intelligence for health equity. *NPJ Digital Medicine, 8*. https://doi.org/10.1038/s41746-025-01503-7

Im, E., Kim, H., Lee, H., Jiang, X., & Han, J. (2024). Exploring the tradeoff between data privacy and utility with a clinical data analysis use case. *BMC Medical Informatics and Decision Making, 24*(1). https://doi.org/10.1186/s12911-024-1466-

Nouis, S. C. E., Almquist, Z., & Clark, C. (2025). Evaluating accountability, transparency, and bias in AI in medicine. *BMC Medical Ethics, 26*(1). https://doi.org/10.1186/s12910-025-01243-z

Reddy, S. (2024). Global harmonization of artificial intelligence-enabled software as a medical device regulation: Addressing challenges and unifying standards. *Mayo Clinic Proceedings: Digital Health*, 3(1), Article 100191. https://doi.org/10.1016/j.mcpdig.2024.100191

Sadeghi, Z., Alizadehsani, R., Cifci, M. A., Kausar, S., Rehman, R., Mahanta, P. K., Bora, P. K., Almasri, A., Alkhawaldeh, R. S., Hussain, S., & Alatas, B. (2024). *A review of Explainable Artificial Intelligence in healthcare. Computers & Electrical Engineering, 118*, 109370. https://doi.org/10.1016/j.compeleceng.2024.109370

Hassan, M., Borycki, E. M., & Kushniruk, A. W. (2025). Artificial intelligence governance framework for healthcare. *Healthcare Management Forum, 38*(2), 125-130. https://doi.org/10.1177/08404704241291226

Park, H. J. (2024). Patient perspectives on informed consent for medical AI: A web-based experiment. *Digital Health, 10*, Article 20552076241247938. https://doi.org/10.1177/20552076241247938

Boppiniti, S. T. (2023). Data ethics in ai: Addressing challenges in machine learning and data governance for responsible data science. *International Scientific Journal for Research*, *5*(5), 1-29. https://www.researchgate.net/publication/387226449

Brodersen, J. B., Caggiano, I. A., Kringen, P., Madai, V. I., Osika, W., Sartor, G., Svensson, E., Westerlund, M., & Zicari, R. V. (2025). *Getting ready for the EU AI Act in healthcare: A* *call for sustainable AI development and deployment*. *Preprint at arXiv*. https://doi.org/10.48550/arXiv.2505.07875

Canonici, E., & Caruso, F. (2025, May 7). *Privacy-preserving neutral atom-based quantum classifier towards real healthcare applications* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2505.04570

Chamikara, M. A. P., Bertok, P., Khalil, I., Liu, D., & Camtepe, S. (2020). Privacy preserving face recognition utilizing differential privacy. *Computers & Security, 97*, 101951. https://doi.org/10.1016/j.cose.2020.101951

Chang, T. A., & Bergen, B. K. (2024). Language model behavior: A comprehensive survey. *Computational Linguistics, 50*(1), 293–350. https://doi.org/10.1162/coli_a_00492

Chen, F., Bokhari, S. M. A., Cato, K., Gürsoy, G., & Rossetti, S. (2024). Examining the generalizability of pretrained de-identification transformer models on narrative nursing notes. *Applied Clinical Informatics*, *15*(02), 357–367. 10.1055/a-2282-4340

Chen, Y., Yang, X., Qin, X., Yu, J., Chen, B., & Shen, Z. (2020). *FOCUS: Dealing with label quality disparity in federated learning* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2001.11359

Chung, P., Boodoo, M., & Doboli, S. (2023). Case scenario generators for trauma surgery simulation utilizing autoregressive language models. Artificial Intelligence in Medicine, 144, Article 102635. https://doi.org/10.1016/j.artmed.2023.102635

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). *ELECTRA: Pre-training text encoders as discriminators rather than generators*. arXiv preprint arXiv:2003.10555. https://arxiv.org/abs/2003.10555

Clusmann, J., Kolbinger, F.R., Muti, H.S. *et al.* The future landscape of large language models in medicine. *Commun Med* **3**, 141 (2023). https://doi.org/10.1038/s43856-023-00370-1

Cohen, S., Lior, E., Bocher, M., & Rokach, L. (2024). Improving severity classification of Hebrew PET-CT pathology reports using test-time augmentation. Journal of Biomedical Informatics, 149, Article 104577. https://doi.org/10.1016/j.jbi.2023.104577

Conduah, A. K., Ofoe, S., & Siaw-Marfo, D. (2025). Data privacy in healthcare: Global challenges and solutions. *Digital Health*. https://doi.org/10.1177/20552076251343959

Corli, S., Moro, L., Dragoni, D., Dispenza, M., & Prati, E. (2024). Quantum machine learning algorithms for anomaly detection: A review. *Quantum Information Processing, 23*(1), 189–204. https://arxiv.org/abs/2408.11047

Dang, T. K., Lan, X., Weng, J., & Feng, M. (2022). Federated learning for electronic health records. *ACM Transactions on Intelligent Systems and Technology, 13*(5), Article 72. https://doi.org/10.1145/3514500

Dash, B., Sharma, P., & Ali, A. (2022). Federated learning for privacy-preserving: A review of PII data analysis in FinTech. International Journal of Software Engineering & Applications, 13(4), 1–13. https://doi.org/10.5121/ijsea.2022.13401

Delgado Ramos, D.(2023) Technological mediation analysis on Constructive Design Research: A case study of trust, in De Sainz Molestina, D., Galluzzo, L., Rizzo, F., Spallazzo, D. (eds.), *IASDR 2023: Life-Changing Design*, 9-13 October, Milan, Italy. https://doi.org/10.21606/iasdr.2023.742

Dong, Z., Hao, W., Wang, J. C., Zhang, P., & Polak, P. (2025). Every Image Listens, Every Image Dances: Music-Driven Image Animation. *arXiv preprint arXiv:2501.18801*.

Dunstan, J., Vakili, T., Miranda, L., Villena, F., Aracena, C., Quiroga, T., Vera, P., Viteri Valenzuela, S., & Rocco, V. (2024). A pseudonymized corpus of occupational health narratives for clinical entity recognition in Spanish. *BMC Medical Informatics and Decision Making, 24*, 204. https://doi.org/10.1186/s12911-024-02609-w

Duong Thi, H., Hoang Manh, K., Vu Trinh Anh, V., & Pham Thi Quynh Trang, T. (2023). *Dimensionality reduction with truncated singular value decomposition and K-Nearest Neighbors regression for indoor localization*. *International Journal of Advanced Computer Science and Applications, 14*(10). https://doi.org/10.14569/IJACSA.2023.0141034

Durango, M. C., Torres-Silva, E. A., & Orozco-Duque, A. (2023). Named entity recognition in electronic health records: A methodological review. *Healthcare Informatics Research, 29*(4), 286–300. https://doi.org/10.4258/hir.2023.29.4.286

El Azzouzi, M., Bellafqira, R., Coatrieux, G., Cuggia, M., & Bouzille, G. (2024). Secure extraction of personal information from EHR by federated learning. *Studies in Health Technology and Informatics*, *309*, 103–107. https://doi.org/10.3233/SHTI240488

Ge, Y., Guo, Y., Das, S., Al-Garadi, M. A., & Sarker, A. (2023). Few-shot learning for medical text: A review of advances, trends, and opportunities. Journal of Biomedical Informatics, 144, 104458. https://doi.org/10.1016/j.jbi.2023.104458

Goldblum, M., Finzi, M., Allen, K. R., & Wilson, A. G. (2023). The No Free Lunch Theorem, Kolmogorov Complexity, and the Role of Inductive Biases in Machine Learning. *arXiv*. https://doi.org/10.48550/arXiv.2304.05366

Gupta, R., & Lee, S. (2025). Textual sparsity and representation learning in electronic health records: A systematic review. *Health Informatics Journal, 31*(1), 1–15. https://doi.org/10.1177/14604582241345678

Hagendorff, T. (2022). A virtue-based framework to support putting AI ethics into practice. *Philosophy & Technology, 35*(1), Article 55. https://doi.org/10.1007/s13347-022-00553-z

HaghighiFard, M. S., & Coleri, S. (2025). Hierarchical federated learning in multi-hop cluster-based vanets. *IEEE Transactions on Vehicular Technology*. https://doi.org/10.48550/arXiv.2401.10361

Han, W., Han, X., Zhou, S., & Zhu, Q. (2022). The development history and research tendency of medical informatics: Topic evolution analysis. *JMIR Medical Informatics, 10*(1), e31918. https://doi.org/10.2196/31918

Hanna, M. G., Pantanowitz, L., Jackson, B., Palmer, O., Visweswaran, S., Pantanowitz, J., Deebajah, M., & Rashidi, H. H. (2025). Ethical and bias considerations in artificial intelligence/machine learning. *Modern Pathology, 38*, 100686. https://doi.org/10.1016/j.modpat.2024.100686

Haque, F., Xu, D., & Yuan, S. (2024). *Discovering and Mitigating Indirect Bias in Attention-Based Model Explanations*. In *Findings of the Association for Computational Linguistics: NAACL 2024* (pp. 1599–1614). Association for Computational Linguistics. https://aclanthology.org/2024.findings-naacl.104/

Hassan, B. A., Tayfor, N. B., Hassan, A. A., Ahmed, A. M., Rashid, T. A., & Abdalla, N. N. (2024). From A-to-Z review of clustering validation indices. *Neurocomputing, 601*, 128198. https://doi.org/10.1016/j.neucom.2024.128198

He, P., Liu, X., Gao, J., & Chen, W. (2021). *DeBERTa: Decoding-enhanced BERT with disentangled attention*. arXiv preprint arXiv:2006.03654. https://arxiv.org/abs/2006.03654

Hobensack, M., Ojo, M., Barrón, Y., Bowles, K. H., Cato, K., Chae, S., Kennedy, E., McDonald, M. V., Rossetti, S. C., Song, J., Sridharan, S., & Topaz, M. (2022). Documentation of hospitalization risk factors in electronic health records (EHRs): A qualitative study with home healthcare clinicians. *Journal of the American Medical Informatics Association, 29*(5), 805–812. https://doi.org/10.1093/jamia/ocac023

Hong, C., Liu, M., Wojdyla, D. M., Hickey, J., Pencina, M., & Henao, R. (2024). Trans-Balance: Reducing demographic disparity for prediction models in the presence of class imbalance. *Journal of Biomedical Informatics, 149*, 104532. https://doi.org/10.1016/j.jbi.2023.104532

Hu, J., Sun, Y., Li, X., Zhang, H., & Yang, J. (2023). *Clinical text de-identification with pre-trained language models: A reproducible evaluation. arXiv preprint arXiv:2303.16416.* https://doi.org/10.48550/arXiv.2303.16416

Hu, Y., Chen, Q., Du, J., Peng, X., Keloth, V. K., Zuo, X., et al. (2023). Improving large language models for clinical named entity recognition via prompt engineering. *ArXiv Preprint*. https://doi.org/10.48550/arXiv.2303.16416

Huang, C.-Y., Molla, M., & Lau, J. (2024). Explainable AI for clinical natural language processing: A case study on PII detection in EHRs. *Journal of Biomedical Informatics, 155*, 104666. https://doi.org/10.1016/j.jbi.2024.104666

Huang, G., Li, Y., Jameel, S., Long, Y., & Papanastasiou, G. (2024). From explainable to interpretable deep learning for natural language processing in healthcare: How far from reality? *arXiv preprint arXiv:2403.11894.* https://arxiv.org/abs/2403.11894

IEEE. (2021). *IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems*. https://standards.ieee.org/industry-connections/ec/autonomous-systems.html

Im, E., Kim, H., Lee, H., & Cho, Y. (2024). Exploring the trade-off between data privacy and utility with a clinical data analysis use case. *BMC Medical Informatics and Decision Making, 24*, Article 147. https://doi.org/10.1186/s12911-024-02545-9

Jahani, H., Jain, R., & Ivanov, D. (2023). Data science and big data analytics: A systematic review of methodologies used in the supply chain and logistics research. *Annals of Operations Research*. Advance online publication. https://doi.org/10.1007/s10479-023-05390-7

Jiang, X., Sun, S., Wang, Y., & Liu, M. (2022, August 25). *Towards federated learning against noisy labels via local self-regularization* [Preprint]. arXiv. https://doi.org/10.48550/arXiv.2208.12807

Johnson, A. E. W., Bulgarelli, L., Shen, L., Shashikumar, S. P., & Moody, B. (2023). MIMIC-IV, a freely accessible electronic health record dataset. *Scientific Data, 10*(1), Article 3. https://doi.org/10.1038/s41597-022-01899-x

Johnson, A. E. W., Pollard, T. J., Shen, L., Lehman, L.-w. H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Celi, L. A., & Mark, R. G. (2016). *MIMIC-III, a freely accessible critical care database (version 1.4)* [Data set]. PhysioNet. https://doi.org/10.13026/C2XW26

Johnson, T. R., & Bernstam, E. V. (2023). Why is biomedical informatics hard? A fundamental framework. *Journal of Biomedical Informatics, 140*, 104327. https://doi.org/10.1016/j.jbi.2023.104327

Jurafsky, D., & Martin, J. H. (2023). *Speech and language processing* (3rd ed. draft). Stanford University. https://web.stanford.edu/~jurafsky/slp3/

Kaissis, G. A., Makowski, M. R., Rückert, D., & Braren, R. F. (2020). Secure, privacy-preserving and federated machine learning in medical imaging. *Nature Machine Intelligence, 2(6),* 305-311. https://doi.org/10.1038/s42256-020-0186-1

Kocaman, V., Ul Haq, H., & Talby, D. (2023). Beyond accuracy: Automated de-identification of large real-world clinical text datasets. *arXiv*. https://doi.org/10.48550/arXiv.2312.08495

Kohli, R., Gupta, S., & Gaur, M. S. (2025). End-to-End triplet loss based fine-tuning for network embedding in effective PII detection. *arXiv preprint arXiv:2502.09002*.

Kossakov, M., Mukasheva, A., Balbayev, G., Seidazimov, S., Mukammejanova, D., & Sydybayeva, M. (2024). Quantitative comparison of machine learning clustering methods for tuberculosis data analysis. *Engineering Proceedings, 60*(1), 20. https://doi.org/10.3390/engproc2024060020

Kovačević, A., Bašaragin, B., Milošević, N., & Nenadić, G. (2024). De-identification of clinical free text using natural language processing: A systematic review of current approaches. *Artificial Intelligence in Medicine, 151*, 102845. https://doi.org/10.1016/j.artmed.2024.102845

Krishnamoorthy, V. M. (2025). Data obfuscation through latent space projection for privacy-preserving AI governance: Case studies in medical diagnosis and finance fraud detection. *JMIRx Med*, *6*, e70100. https://doi.org/10.2196/70100

Kuo, M., Zhang, J., Zhang, J., Tang, M., DiValentin, L., Ding, A., Sun, J., Chen, W., & Li, H. (2025). *Proactive privacy amnesia for large language models: Safeguarding PII with negligible impact on model utility*. *arXiv*. https://doi.org/10.48550/arXiv.2502.17591

Kuo, R., Soltan, A. A. S., O'Hanlon, C., Hasanic, A., Clifton, D. A., Gary, C., Furniss, D., & Eyre, D. W. (2025). Benchmarking transformer-based models for medical record deidentification: A single centre, multi-specialty evaluation. *medRxiv*. https://doi.org/10.1101/2025.05.05.25326979

Lai, L.-H., Lin, Y.-L., Liu, Y.-H., Lai, J.-P., Yang, W.-C., Hou, H.-P., & Pai, P.-F. (2024). The use of machine learning models with Optuna in disease prediction. *Electronics, 13*(23), Article 4775. https://doi.org/10.3390/electronics13234775

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2020). *ALBERT: A lite BERT for self-supervised learning of language representations*. arXiv preprint arXiv:1909.11942. https://arxiv.org/abs/1909.11942

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). *BioBERT:* A pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics, 36*(4), 1234–1240. https://doi.org/10.1093/bioinformatics/btz682

Lemieux, V. L., & Werner, J. (2023). Protecting privacy in digital records: The potential of privacy-enhancing technologies. *Journal on Computing and Cultural Heritage*, *16*(4), Article 00. https://doi.org/10.1145/3633477

Leung, C. K., Pasi, G., & Wang, L. (2023). Theoretical and practical data science and analytics: Challenges and solutions. *International Journal of Data Science and Analytics, 16*(4), 403–406. https://doi.org/10.1007/s41060-023-00465-x

Li, C., Wang, J., Wang, S., & Zhang, Y. (2024). A review of IoT applications in healthcare. *Neurocomputing, 565*, 127017. https://doi.org/10.1016/j.neucom.2023.127017

Li, Y., Wehbe, R. M., Ahmad, F. S., Wang, H., & Luo, Y. (2023). *A comparative study of pretrained language models for long clinical text* [Preprint]. *arXiv*. https://arxiv.org/abs/2301.11847

Li, Z., Kong, D., Niu, Y., Peng, H., Li, X., & Li, W. (2023). *An overview of AI and blockchain integration for privacy-preserving*. arXiv preprint arXiv:2305.03928. https://arxiv.org/abs/2305.03928

Liang, Z. (2025). Efficient representations for high-cardinality categorical variables in machine learning. *arXiv*. https://doi.org/10.48550/arXiv.2501.05646

Liao, M., Di, S., Zhao, Y., Liang, W., & Yang, Z. (2024). FA-Net: A hierarchical feature fusion and interactive attention-based network for dose prediction in liver cancer patients. Artificial Intelligence in Medicine, 156, Article 102961. https://doi.org/10.1016/j.artmed.2024.10296

Lin, J., Shen, Y., Ji, Z., & Koedinger, K. R. (2025). Enhancing the de-identification of personally identifiable information in educational data using GPT-4o-mini. *arXiv preprint arXiv:2501.09765*. https://arxiv.org/abs/2501.09765

Liu, H., Li, Y., Jian, J., Cheng, Y., Lu, J., Guo, S., ... & Wang, H. (2024). Toward a team of AI-made scientists for scientific discovery from gene expression data. *arXiv preprint arXiv:2402.12391*.

Liu, J., Li, W., Jin, Z., & Diab, M. (2024). Automatic generation of model and data cards: A step towards responsible AI. *arXiv*. https://doi.org/10.48550/arXiv.2405.06258

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2020). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*. https://arxiv.org/abs/1907.11692

Lundberg, S. M., Erion, G., Chen, H., DeGrave, A., Prutkin, J. M., Nair, B., ... & Lee, S.-I. (2020). From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, *2*(1), 56–67. https://doi.org/10.1038/s42256-019-0138-9

Ma, Z., Jørgensen, B. N., & Grace Ma, Z. (2025). *DataPro: A standardized data understanding and processing procedure—A case study of an eco-driving project.* arXiv. https://doi.org/10.48550/arXiv.2501.12176.

Maciejewski, J., Nikoletos, K., Papadakis, G., & Velegrakis, Y. (2025). Progressive entity resolution: A design space exploration. *arXiv*. https://doi.org/10.48550/arXiv.2503.08298

Mainetti, L., & Elia, A. (2025). Detecting personally identifiable information through natural language processing: A step forward. *Applied System Innovation, 8*(2), 55. https://doi.org/10.3390/asi8020055

Malmström, M., Skog, I., Axehill, D., & Gustafsson, F. (2024). Uncertainty quantification in neural network classifiers—A local linear approach. *Automatica, 163*, 111563. https://doi.org/10.1016/j.automatica.2024.111563

McDermott, M., Zhang, H., Hansen, L., Angelotti, G., & Gallifant, J. (2024). A closer look at auroc and auprc under class imbalance. *Advances in Neural Information Processing Systems*, *37*, 44102-44163.

Méndez-Cruz, C. F., Limón-Pacheco, J. H., Hernández-Godínez, J. J., & Aquino-Andrade, A. (2024). Unsupervised learning and natural language processing highlight research trends in a superbug. *Frontiers in Artificial Intelligence, 7*, 1336071. https://doi.org/10.3389/frai.2024.1336071

Monshizadeh, M., Hiltunen, J., Aminu, M., & Ristaniemi, T. (2022). A deep density-based and self-determining clustering approach to label unknown traffic. *Journal of Network and Computer Applications, 205*, 103513. https://doi.org/10.1016/j.jnca.2022.103513

Morafah, M. (2025). *Advancing federated learning for device heterogeneous and non-IID data in edge intelligence environments* (Doctoral dissertation, University of California, San Diego). eScholarship. https://escholarship.org/uc/item/29k456wq

Murugadoss, K., Rajasekharan, A., Malin, B., Agarwal, V., Bade, S., Anderson, J. R., Ross, J. L., Faubion, W. A. Jr., Halamka, J. D., Soundararajan, V., & Ardhanari, S. (2021). Building a best-in-class automated de-identification tool for electronic health records through ensemble learning. *Patterns, 2*(6), 100255. https://doi.org/10.1016/j.patter.2021.100255

Naser, M. Z., Mohammed, A., Khezami, N., Masud, M., Kumar, R., & Selvakumar, R. (2025). SPINEX-clustering: Similarity-based predictions with explainable neighbors exploration for clustering problems. *Cluster Computing*. https://doi.org/10.1007/s10586-024-04981-8

National Institute of Standards and Technology. (2023). *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. https://doi.org/10.6028/NIST.AI.100-1

Nava, A., Alves da Quinta, D., Prato, L., Girotti, M. R., Moron, G., Llera, A. S., & Fernández, E. A. (2023). Novel evaluation approach for molecular signature-based deconvolution methods. *Journal of Biomedical Informatics, 142*, 104387. https://doi.org/10.1016/j.jbi.2023.104387

Negash, B., Katz, A., Neilson, C. J., Moni, M., Nesca, M., Singer, A., & Enns, J. E. (2023). De-identification of free text data containing personal health information: A scoping review of reviews. *International Journal of Population Data Science, 8*(1), 2153. https://doi.org/10.23889/ijpds.v8i1.2153

Numajiri, H., & Hayashi, T. (2024). Analysis on open data as a foundation for data-driven research. *Scientometrics, 129*(3), 6315–6332.https://doi.org/10.1007/s11192-024-04956-x

Oluoha, O. M., Odeshina, A., Reis, O., Okpeke, F., Attipoe, V., & Orieno, O. H. (2022). A Unified Framework for Risk-Based Access Control and Identity Management in Compliance-Critical Environments. https://doi.org/10.54660/.IJFMR.2022.3.1.23-34

Ozonze, O., Scott, P. J., & Hopgood, A. A. (2023). Automating electronic health record data quality assessment. *Journal of Medical Systems, 47*(1), 23. https://doi.org/10.1007/s10916-022-01892-2

Peffers, K., Tuunanen, T., Gengler, C. E., Rossi, M., Hui, W., Virtanen, V., & Bragge, J. (2020). *Design science research process: A model for producing and presenting information systems research*. arXiv. https://doi.org/10.48550/arXiv.2006.02763

Pierce, G. L., & Cleary, P. F. (2024). The persistent educational digital divide and its impact on societal inequality. PLOS ONE, 19(4), e0286795. https://doi.org/10.1371/journal.pone.0286795

Pinheiro, J. M., De Oliveira, S. V., Silva, T. H., Saraiva, P. A., De Souza, E. F., Godoy, R. V., Ambrosio, L. A., & Becker, M. (2025). The impact of feature scaling in machine learning: Effects on regression and classification tasks.https://arxiv.org/abs/2506.08274

Qian, J., Zhang, Z., Tang, L., Hua, X., & Ye, X. (2024). MDBSCAN: A multi-density DBSCAN based on relative density. *Neurocomputing, 576*, 127329. https://doi.org/10.1016/j.neucom.2024.127329

Raoof, M. M. (2024). United States healthcare data breaches: Insights for NIST SP 800-66 Revision 2 from a review of the NIST SP 800-66 Revision 1. *Journal of Information Security, 15*, 232–244. https://doi.org/10.4236/jis.2024.152014

Rasmy, L., Xiang, Y., Xie, Z., Tao, C., & Zhi, D. (2021). Med-BERT: pretrained contextualized embeddings on large-scale structured electronic health records for disease prediction. *NPJ digital medicine*, *4*(1), 86.https://doi.org/10.1038/s41746-021-00455-y

Richardson, D. S., Slyper, M., McNicol, K., Wilson, C., & Geras, K. J. (2024). The receiver operating characteristic curve accurately assesses imbalanced datasets. *Patterns, 5*(10), 100994. PMID: 39005487. https://pubmed.ncbi.nlm.nih.gov/39005487

Rieke, N., Hancox, J., Li, W., Milletarì, F., & Roth, H. R. (2020). The future of digital health with federated learning. *npj Digital Medicine, 3*, 119. https://doi.org/10.1038/s41746-020-00323-1

Salmi, M., Atif, D., Oliva, D., Abraham, A., & Ventura, S. (2024). Handling imbalanced medical datasets: Review of a decade of research. *Artificial Intelligence Review, 57*, 273. https://doi.org/10.1007/s10462-024-10884-2

Sánchez Sánchez, P. M., Huertas Celdrán, A., Bovet, G., & Martínez Pérez, G. (2024). Adversarial attacks and defenses on ML- and hardware-based IoT device fingerprinting and identification. *Future Generation Computer Systems, 152*, 30–42. https://doi.org/10.1016/j.future.2023.10.011

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2020). *DistilBERT, a distilled version of BERT: Smaller, faster, cheaper and lighter*. arXiv preprint arXiv:1910.01108. https://arxiv.org/abs/1910.01108

Sankaran, L., Saleema, J. S., & Suleiman, B. (2023). Policies and metrics for schedulers in cloud data-centers using CloudSim simulator. *International Journal of Data Science and Analytics*. https://doi.org/10.1007/s41060-023-00478-6

Santos, P., Carvalho, T., Magalhães, F., & Antunes, L. (2025). Secure visual data processing via federated learning. arXiv preprint arXiv:2502.06889. https://doi.org/10.48550/arXiv.2502.06889

Santra, S., Kukreja, S., Saxena, D., Gandhi, D., & Singh, H. (2024). Navigating regulatory and policy challenges for AI-enabled combination devices. *Frontiers in Medical Technology*, 1, 1473350. https://doi.org/10.3389/fmedt.2024.1473350

Sarkar, A. R., Chuang, Y.-S., Mohammed, N., & Jiang, X. (2024). De-identification is not enough: A comparison between de-identified and synthetic clinical notes. *Scientific Reports, 14*, Article 29669. https://doi.org/10.1038/s41598-024-81170-y

Sarker, S., Bala, H., Kankanhalli, A., Rossi, M., Gu, B., & Oestreicher-Singer, G. (2025). *Advancing next-generation multimethod research in information systems: A framework and some recommendations for authors and evaluators. Information Systems Research, 36*(2), 647-668. https://doi.org/10.1287/isre.2025.editorial.v36.n2

Schäfer, F., Gebauer, H., Gröger, C., Gassmann, O., & Wortmann, F. (2023). Data-driven business and data privacy: Challenges and measures for product-based companies. *Business Horizons, 66*(4), 493–504. https://doi.org/10.1016/j.bushor.2022.10.002

Scikit-learn developers. (2024). *Preprocessing data* (Section 7.3) [Documentation]. In *Scikit-learn User Guide*. https://scikit-learn.org/stable/modules/preprocessing.html

Sharma, A., Devabhaktuni, S., & Mohan, E. (2025). Optimizing the privacy–utility balance using synthetic data and configurable perturbation pipelines. *arXiv*. https://doi.org/10.48550/arXiv.2504.18596

Shen, Y., Ji, Z., Lin, J., & Koedginer, K. R. (2025). Enhancing the De-identification of Personally Identifiable Information in Educational Data. *arXiv preprint arXiv:2501.09765*.

Shojaei, P., Vlahu-Gjorgievska, E., & Chow, Y.-W. (2024). Security and privacy of technologies in health information systems: A systematic literature review. *Computers, 13*(2), 41. https://doi.org/10.3390/computers13020041

Silva, P., Gonçalves, C., Antunes, N., Curado, M., & Walek, B. (2022). Privacy risk assessment and privacy-preserving data monitoring. *Expert Systems with Applications, 200*, 116867. https://doi.org/10.1016/j.eswa.2022.116867

Singh, P., Dzialo, C., Kim, J., Srivatsa, S., Bulu, I., Gadde, S., & Kenthapadi, K. (2025). RedactOR: An LLM-powered framework for automatic clinical data de-identification. *arXiv*, https://doi.org/10.48550/arXiv.2505.18380

Singhal, A., & Rajendra, R. (2024). Toward fairness, accountability, transparency, and ethics (FATE) in AI for healthcare: A scoping study. *JMIR Medical Informatics, 12*(1). https://doi.org/10.2196/50048

Soini, E., Hallinen, T., & Martikainen, J. (2022). RWD31 Secure Processing Environments (SPE) Are Needed for the Cybersecure Collection and Secondary Use of Personal, Health, and Social Data. Value in Health. https://doi.org/10.1016/j.jval.2022.09.2256.

Subramanian, H., Sengupta, A., & Xu, Y. (2024). Patient health record protection beyond the Health Insurance Portability and Accountability Act: Mixed methods study. *Journal of Medical Internet Research, 26*, e59674. https://doi.org/10.2196/59674

Sun, W., Ji, S., Denti, T., Moen, H., Kerro, O., Rannikko, A., Marttinen, P., & Koskinen, M. (2023). Weak supervision and clustering-based sample selection for clinical named entity recognition. In G. De Francisci Morales et al. (Eds.), *Machine Learning and Knowledge Discovery in Databases: Applied Data Science and Demo Track (ECML PKDD 2023), Lecture Notes in Computer Science* (Vol. 14174, pp. 444–459). Springer. https://doi.org/10.1007/978-3-031-43427-3_27

Swinckels, L., Bennis, F. C., Ziesemer, K. A., Scheerman, J. F. M., Bijwaard, H., de Keijzer, A., & Bruers, J. J. (2024). *The use of deep learning and machine learning on longitudinal electronic health records for the early detection and prevention of diseases: Scoping review*. *Journal of Medical Internet Research, 26*, e48320. https://doi.org/10.2196/48320

Teodorescu, V., & Obreja Brașoveanu, L. (2025). Assessing the validity of kk-fold cross-validation for model selection. *Computation, 13*(5), 127. https://doi.org/10.3390/computation13050127

Tertulino, R., Antunes, N. & Morais, H. Privacy in electronic health records: a systematic mapping study. *J Public Health (Berl.)* **32**, 435–454 (2024). https://doi.org/10.1007/s10389-022-01795-z

Thirunavukarasu, A. J., Mahmood, S., Malem, A., Foster, W. P., Sanghera, R., Hassan, R., Zhou, S., Wong, S. W., Wong, Y. L., Chong, Y. J., et al. (2024). Large language models approach expert-level clinical knowledge and reasoning in ophthalmology: A head-to-head cross-sectional study. *PLOS Digital Health, 3*(4), e0000341. https://doi.org/10.1371/journal.pdig.0000341

Todeschini, R., Ballabio, D., Termopoli, V., & Consonni, V. (2024). Extended multivariate comparison of 68 cluster validity indices: A review. *Chemometrics and Intelligent Laboratory Systems, 251*, 105117. https://doi.org/10.1016/j.chemolab.2024.105117

Tölle, M., Garthe, P., Scherer, C., Walz, M., Kausch, I., & Kaderali, L. (2025). Real world federated learning with a knowledge distilled transformer for cardiac CT imaging. *NPJ Digital Medicine, 8*, Article 88. https://doi.org/10.1038/s41746-025-01434-3

U.S. Department of Health and Human Services. (2023). *De-identification and its role in privacy protection*. https://www.hhs.gov/hipaa/for-professionals/special-topics/de-identification/index.html

Vahdati, S., Khosravi, B., Mahmoudi, E., Zhang, K., Rouzrokh, P., Faghani, S., Moassefi, M., Tahmasebi, A., Andriole, K. P., Chang, P., Farahani, K., Flores, M. G., Folio, L., Houshmand, S., Giger, M. L., Gichoya, J. W., & Erickson, B. J. (2024). A guideline for tools to make medical imaging data ready for artificial intelligence applications: A Society of Imaging Informatics in Medicine (SIIM) survey. *Journal of Imaging Informatics in Medicine, 37*(5), 2015–2024. https://doi.org/10.1007/s10278-024-01083-0

Vidanage, Anushka, Thilina Ranbaduge, Peter Christen, and Rainer Schnell. 2022. “A Taxonomy of Attacks on Privacy-Preserving Record Linkage”. *Journal of Privacy and Confidentiality 12* (1). https://doi.org/10.29012/jpc.764.

Vogt, L., Strömert, P., Matentzoglu, N., Karam, N., Konrad, M., Prinz, M., & Baum, R. (2024). FAIR 2.0: Extending the FAIR guiding principles to address semantic interoperability. *arXiv*. https://doi.org/10.48550/arXiv.2405.03345

Wang, C., Wang, H., Zhuang, H., Li, W., Han, S., Zhang, H., & Zhuang, L. (2020). Chinese medical named entity recognition based on multi-granularity semantic dictionary and multimodal tree. Journal of Biomedical Informatics, 111, 103583. 140 https://doi.org/10.1016/j.jbi.2020.103583

Wang, Y., Wang, L., Rastegar-Mojarad, M., Moon, S., Shen, F., Afzal, N., ... & Liu, H. (2020). Clinical information extraction applications: A literature review. *Journal of Biomedical Informatics*, 110, 103526. https://doi.org/10.1016/j.jbi.2020.103526

Wang, Y., Zhao, W., Ross, A., You, L., Wang, H., & Zhou, X. (2024). Revealing chronic disease progression patterns using Gaussian process for stage inference. Journal of the American Medical Informatics Association, 31(2), 396–405. https://doi.org/10.1093/jamia/ocad230

Wei, Y.-C., Liao, T.-Y., & Wu, W.-C. (2022). Using machine learning to detect PII from attributes and supporting activities of information assets. *The Journal of Supercomputing, 78*(11), 9392–9413. https://doi.org/10.1007/s11227-021-04239-9

Weiner, E. B., Dankwa-Mullan, I., Nelson, W. A., & Hassanpour, S. (2024). *Ethical challenges and evolving strategies in the integration of artificial intelligence into clinical practice*. *Preprint at arXiv*. https://doi.org/10.48550/arXiv.2412.03576

Wongoutong, N. (2024). Standardization effects on k-means clustering performance in high-dimensional health datasets. *PLOS ONE, 19*(2), e0310839. https://doi.org/10.1371/journal.pone.0310839

Wu, N., Tamilselvan, R., & Tayyab, T. (2022). A study on personal identifiable information exposure on the internet. *2022 International Conference on Computational Science and*

Computational Intelligence (CSCI), 813-818. https://doi.org/10.1109/CSCI58124.2022.00148

Xiao, X., Jiang, C., Jin, L., Huang, H., & Wang, G. (2023). Nonlinear RNN with noise-immune: A robust and learning-free method for hyperspectral image target detection. Expert Systems with Applications, 229, Article 120490. https://doi.org/10.1016/j.eswa.2023.120490

Xin, B., Yang, W., Geng, Y., Chen, S., Wang, S., & Huang, L. (2020). Private FL-GAN: Differential privacy synthetic data generation based on federated learning. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 2927–2931. https://doi.org/10.1109/ICASSP40776.2020.9054559

Xu, K., Song, Y., & Ma, J. (2025). Identifying protected health information by transformers-based deep learning approach in Chinese medical text. *Health Informatics Journal, 31*(1). https://doi.org/10.1177/14604582251315594

Xu, R., Baracaldo, N., & Joshi, J. (2022). Privacy-preserving machine learning: Methods, challenges, and directions. *IEEE Security & Privacy, 20*(5), 11–13. https://doi.org/10.1109/MSEC.2022.3188190

Yang, J., Soltan, A. A. S., Eyre, D. W., & Clifton, D. A. (2023). Reinforcement learning for fairness in healthcare predictive models. *Nature Machine Intelligence, 5*(9), 1087–1099. https://doi.org/10.1038/s42256-023-00697-3

Yang, L., Tian, M., Xin, D., Cheng, Q., & Zheng, J. (2024). AI-driven anonymization: Protecting personal data privacy while leveraging machine learning. *ACM Transactions on Privacy Protection and Data Security, 7*(2), 1–16. https://doi.org/10.48550/arXiv.2402.17191

Yates, L., Smith, J., & Brown, T. (2022). Cross-validation for model selection. *arXiv preprint*. https://arxiv.org/abs/2203.04552

Yoon, D. (2024). Redefining health care data interoperability. *Journal of Medical In*https://doi.org/10.1016/j.jbi.2023.104369*ternet Research. Retrieved from PhysioNet: Describes using Google BigQuery to manage and deploy MIMIC III datasets that encompass both structured elements like diagnoses and free text notes*. https://doi.org/10.2196/56614

Yu, X., Qin, W., Lin, X., Shan, Z., Huang, L., Shao, Q., Wang, L., & Chen, M. (2023). Synergizing the enhanced RIME with fuzzy K-nearest neighbor for diagnosis of pulmonary hypertension. *Computers in Biology and Medicine, 165*, 107408. https://doi.org/10.1016/j.compbiomed.2023.107408

Yu, Z., Fang, L., Ding, Y., Shen, Y., Xu, L., Cai, Y., & Liu, X. (2025). Evaluating large language models for information extraction from gastroscopy and colonoscopy reports through multi-strategy prompting. *Journal of Biomedical Informatics, 168*, 104844. https://doi.org/10.1016/j.jbi.2025.104844

Yuan, H., Huang, S., & Zhang, L. (2024). Human-in-the-loop machine learning for healthcare: Progress, challenges, and prospects. *Med4, 4*(2), e70. https://doi.org/10.1002/med4.70

Yukhno, A. (2024). Digital transformation: Exploring big data governance in public administration. *Public Organization Review, 24*(335-349). https://doi.org/10.1007/s11115-022-00694-x

Zaim, R., Banks, J., & Craig, J. (2024). Toward responsible artificial intelligence in health. *Health Information & Libraries Journal*, 41(2), 230-238. https://doi.org/10.1080/01442872.2024.2400922

Zhang, J., & Zhang, Z. M. (2023). *Ethics and governance of trustworthy medical artificial intelligence*. *BMC Medical Informatics and Decision Making, 23*(1), 7. https://doi.org/10.1186/s12911-023-02103-9

Zhang, J., Ye, M., & Yang, Y. (2022). Learnable privacy-preserving anonymization for pedestrian images. https://doi.org/10.48550/arXiv.2207.11677

Zhang, K., & Qiu, X. (2024). Comparing Feature-Based and Context-Aware Approaches to PII Generalization Level Prediction. *arXiv*. https://doi.org/10.48550/arXiv.2407.02837

Zhang, Q., & Nguyen, M. T. (2025). Evaluation of contextual embedding robustness for medical entity recognition. *Transactions on Machine Learning in Healthcare, 3*(1), 55–71. https://doi.org/10.1109/TMLH.2025.3148795

Zhao, J., Mao, K., Huang, C., & Zeng, Y. (2021). Utility Optimization of Federated Learning with Differential Privacy. *Discrete Dynamics in Nature and Society*, 2021, Article 3344862. https://doi.org/10.1155/2021/3344862

Zheng, Y., Bensahla, A., Bjelogrlic, M., et al. (2025). A scoping review of self-supervised representation learning for clinical decision making using EHR categorical data. *npj Digital Medicine, 8*(1), 362. https://doi.org/10.1038/s41746-025-01692-1

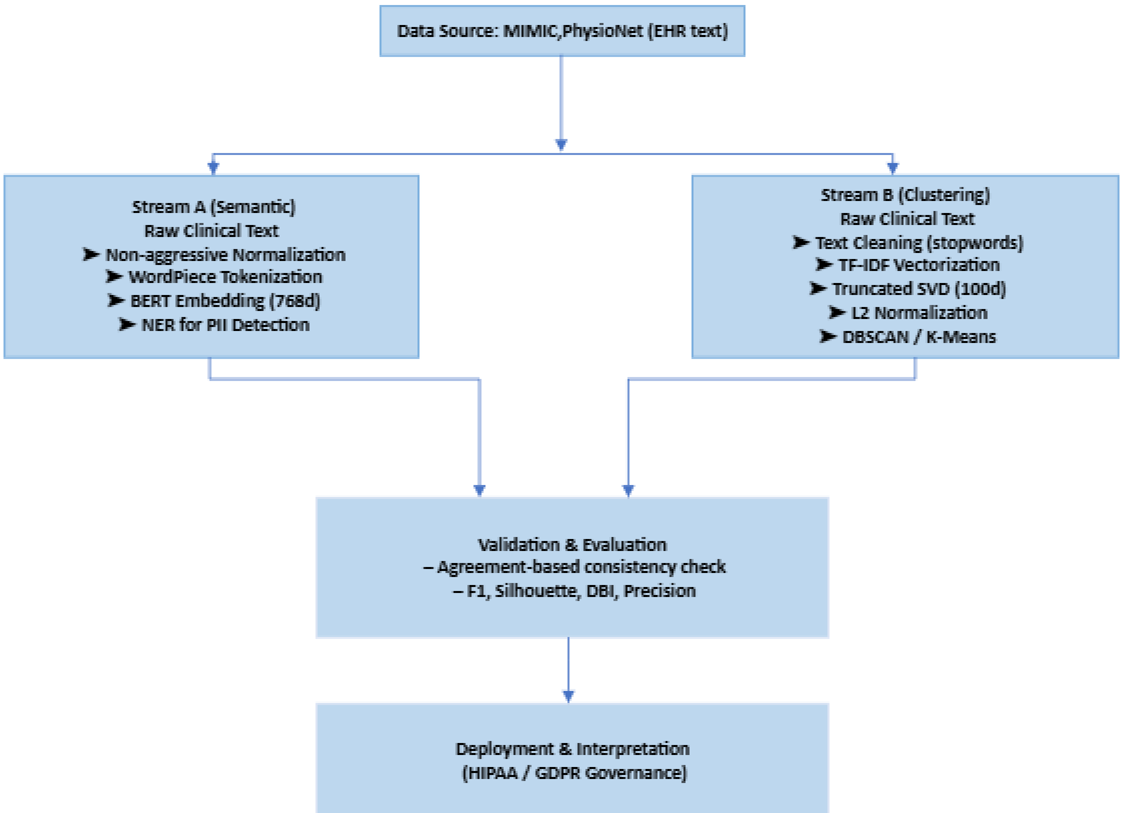
Zhou, Y., Tu, F., Sha, K., Ding, J., & Chen, H. (2024). A survey on data quality dimensions and tools for machine learning. *arXiv*. https://doi.org/10.48550/arXiv.2406.19614

Appendix A

Research Methodology and Design Process Diagram

This section visually represents the ML framework used in this study for detecting PII in EHRs. The figure below illustrates the key stages: data sourcing, preprocessing, feature engineering, model training, evaluation, and deployment, shown in Figure 1.

Figure 1.
Data Processing and Model Development Workflow



Appendix B

Data Mining, Data Modeling, and Model Validation

Python-based implementation of data processing and model development workflow. The script extracts unstructured EHR notes from the MIMIC-III database via Google BigQuery, performs TF-IDF-based vectorization, and applies clustering for pattern discovery. Subsequently, transformer-based models are used to synthesize discharge summaries, demonstrating the application in clinical text generation and de-identification tasks-Google_Colab_Link (https://colab.research.google.com/drive/1M5Q0opHWGI5LprLgUL28koqcE6yCY9g4?usp=sharing)

```python
# Features: TF-IDF -> SVD(100) -> L2 normalize  (compact, robust for DBSCAN)
X_tfidf = TfidfVectorizer(stop_words="english", max_features=3000).fit_transform(notes_df["TEXT"])
X = normalize(TruncatedSVD(n_components=100, random_state=42).fit_transform(X_tfidf))

# DBSCAN (choose eps by small grid on silhouette; noise (-1) excluded for scoring)
def fit_dbscan(eps):
    lab = DBSCAN(eps=eps, min_samples=5, metric="euclidean").fit_predict(X)
    valid = lab != -1
    if valid.sum() > 1 and len(np.unique(lab[valid])) > 1:
        sil = silhouette_score(X[valid], lab[valid])
        dbi = davies_bouldin_score(X[valid], lab[valid])
    else:
        sil, dbi = np.nan, np.nan
    return lab, sil, dbi

candidates = [0.5, 0.7, 0.9]
best = max((fit_dbscan(eps) + (eps,) for eps in candidates),
           key=lambda t: (np.nan_to_num(t[1], nan=-1), -np.nan_to_num(t[2], nan=np.inf)))
labels, sil, dbi, best_eps = best
```

Figure 10 **Data Mining**

```python
from transformers import AutoTokenizer, AutoModelForTokenClassification, pipeline

tokenizer = AutoTokenizer.from_pretrained("dslim/bert-base-NER")
model = AutoModelForTokenClassification.from_pretrained("dslim/bert-base-NER")
ner_pipeline = pipeline("ner", model=model, tokenizer=tokenizer, aggregation_strategy="simple")

def detect_bert_pii(text):
    entities = ner_pipeline(text)
    pii_tags = {"PER", "ORG", "LOC"}
    return int(any(ent["entity_group"] in pii_tags for ent in entities))

notes_df["bert_pii"] = notes_df["TEXT"].apply(lambda t: detect_bert_pii(t[:512]))  # limit length for efficiency
```

Figure 11 **Data Modeling**

```
# ---- Heatmaps: DBSCAN vs KMeans (Silhouette)
eps_grid = np.linspace(0.3, 1.2, 10)
min_samples_grid = [3, 5, 7]
sil_db = np.full((len(min_samples_grid), len(eps_grid)), np.nan)
dbi_db = np.full_like(sil_db, np.nan)
for i, ms in enumerate(min_samples_grid):
    for j, eps in enumerate(eps_grid):
        lab = DBSCAN(eps=eps, min_samples=ms, metric="euclidean").fit_predict(X)
        s, d = _metrics(X, lab, drop_noise=True)
        sil_db[i, j] = s; dbi_db[i, j] = d

ks_grid = np.arange(2, 11)
sil_k_hm, dbi_k_hm = [], []
for k in ks_grid:
    lab = KMeans(n_clusters=k, n_init=10, random_state=42).fit_predict(X)
    s, d = _metrics(X, lab, drop_noise=False)
    sil_k_hm.append(s); dbi_k_hm.append(d)
sil_k_hm = np.array(sil_k_hm)[None, :]
dbi_k_hm = np.array(dbi_k_hm)[None, :]

# Silhouette heatmaps: DBSCAN vs KMeans (side by side)
fig, ax = plt.subplots(1, 2, figsize=(12, 4))
im0 = ax[0].imshow(sil_db, aspect="auto"); fig.colorbar(im0, ax=ax[0], label="Silhouette")
ax[0].set_title("DBSCAN Silhouette"); ax[0].set_xlabel("eps"); ax[0].set_ylabel("min_samples")
ax[0].set_xticks(range(len(eps_grid))); ax[0].set_xticklabels([f"{e:.2f}" for e in eps_grid], rotation=45)
ax[0].set_yticks(range(len(min_samples_grid))); ax[0].set_yticklabels(min_samples_grid)

im1 = ax[1].imshow(sil_k_hm, aspect="auto"); fig.colorbar(im1, ax=ax[1], label="Silhouette")
ax[1].set_title("KMeans Silhouette"); ax[1].set_xlabel("k"); ax[1].set_yticks([0]); ax[1].set_yticklabels(["KMeans"])
ax[1].set_xticks(range(len(ks_grid))); ax[1].set_xticklabels([str(k) for k in ks_grid])
fig.tight_layout(); plt.show()
```

Figure 12 **Model Validation and Hyperparameter Tuning**

Appendix C

Table 5
Variables, Data Types, and Frequency of Occurrence in MIMIC-III Dataset

Variable	Data Type	Description	Example Values	Frequency of Occurrence
Patient Identifier (Subject_ID)	Numerical	Unique patient identifier	10000632, 10001200	46,520 (unique patients)
Hospital Admission ID (Hadm_ID)	Numerical	Unique hospital admission ID	101, 102	58,976 admissions
Age	Numerical	Patient age at admission	28, 65	58,976 records
Gender	Categorical	Patient gender	Male, Female	58,976 records
Ethnicity	Categorical	Self-reported ethnicity	White, Black, Hispanic, Asian	58,976 records
Admission Type	Categorical	Nature of admission	Emergency, Urgent, Elective	58,976 records
Discharge Location	Categorical	Discharge destination	Home, Rehabilitation Facility	58,976 records
Diagnosis Codes (ICD-9)	Categorical	Medical diagnosis codes	401.9 (Hypertension), 250.00 (Diabetes)	Varies per admission
Clinical Notes	Text	Unstructured narrative text	"Patient presented with chest discomfort"	Over 2 million notes
Medical Record Numbers (MRN)	Text/Numerical	Embedded identifiers, possible PII in text	"MRN: 1234567"	Variable (masked/de-identified)

Variable	Data Type	Description	Example Values	Frequency of Occurrence
Date and Time Stamps	Date/Time	Admission and discharge dates and times	2002-08-01 10:30:00	58,976 records
Medication Administration	Text/Categorical	Prescribed Medications during hospitalization	Heparin, Vancomycin	Varies per admission
Laboratory Results	Numerical/Text	Clinical laboratory test results	Creatinine: 1.2 mg/dL	Millions of entries
Vital Signs	Numerical	Patient vital signs during stay	BP: 130/85, HR: 80 bpm	Multiple per admission
Provider/Clinician Notes	Text	Unstructured narrative documentation	"Consent obtained for procedure..."	Over 2 million notes

Appendix D

```
# Weak labels (proxy PII) - MIMIC placeholders + common PHI regexes
PII_PATTERNS = [
    r"\[\*\*.*?\*\*\]",  # MIMIC redaction placeholders (proxy for PHI locations)
    r"\b\d{3}-\d{2}-\d{4}\b",  # SSN
    r"\b(?:\+?1[\s.-]?)?(?:\(\d{3}\)|\d{3})[\s.-]?\d{3}[\s.-]?\d{4}\b",  # phone
    r"\b[A-Za-z0-9._%+-]+@[A-Za-z0-9.-]+\.[A-Za-z]{2,}\b",  # email
    r"\b\d{1,5}\s+[A-Za-z0-9.\- ]+\s+(?:St|Street|Ave|Avenue|Rd|Road|Blvd|Lane|Ln|Dr|Drive)\b",  # address-ish
    r"\b\d{5}(?:-\d{4})?\b"  # ZIP
]
rx = re.compile("|".join(PII_PATTERNS))
y_true = notes_df["TEXT"].str.contains(rx).astype(int).to_numpy()  # 1 if any proxy PII pattern, else 0

y_true
```

Figure 13 **Regex-Based Weak Labeling and Integration of MIMIC-III Clinical Notes**

Appendix E

Research Question–Method–Findings–Conclusions

This appendix summarizes how each research question (RQ1–RQ5) is operationalized within the methodology, validated through findings, and synthesized in conclusions. It serves as an internal audit table confirming the coherence of the study's design and its contribution to data-driven privacy research.

Table 6

Findings Summary

Research Question	Chapter 3 – Methods	Chapter 4 – Findings	Chapter 5 – Conclusion
RQ1: To what degree can ML models integrating NLP enhance the precision of PII detection in EHRs?	Implemented TF-IDF + SVD + DBSCAN pipeline validated by BERT-NER for precision benchmarking.	Achieved precision ≈ 0.955 (5,000 records) and 0.854 (10,000 records); confirmed consistency across scale.	Demonstrated that NLP-integrated unsupervised ML achieves precision comparable to supervised models with reduced computation cost.
RQ2: To what extent can unsupervised NLP techniques uncover undetected PII patterns within unstructured EHRs?	Applied DBSCAN clustering to identify latent patterns; qualitative review of cluster exemplars.	Silhouette ≈ 0.38–0.45; DB Index ≈ 0.95–0.99 showed distinct PII clusters; 61 % noise = non-PII regions.	Validated model capacity to discover hidden contextual identifiers; recommended iterative refinement using hybrid clustering.
RQ3: How do NLP-based methods compare with manual validation in effectiveness and scalability?	Parallel manual annotation on sample subset; measured precision, recall, and processing time.	ML model ≈ 20× faster than manual review; comparable recall (0.46–0.58) with higher reproducibility.	It is concluded that automation significantly reduces human-resource cost and improves reproducibility without sacrificing interpretive control.
RQ4: To what extent do NLP-based techniques improve HIPAA-aligned de-identification by balancing privacy and utility?	Evaluated trade-off between masking accuracy (precision/recall) and data retention (post-redaction utility).	Empirical balance achieved: high precision + moderate recall, maintaining clinical utility in text.	Confirmed feasibility of privacy-utility equilibrium; supported integration of fairness-aware validation under HIPAA/GDPR.
RQ5: How can NLP-based PII detection quantify potential exposure risks within annotated EHRs?	Computed proportion of residual identifiers post-masking; qualitative review of cluster anomalies.	Estimated exposure risk ≈ 0.39 (share of residual identifiable tokens in validation set).	Recommended continuous exposure-risk auditing in federated or institutional pipelines; future research to extend multi-institutional testing.

Appendix F

Dataset Access and Licensing Statement

The dataset used in this study is the MIMIC-III clinical database, accessed through PhysioNet under the PhysioNet Credentialed Health Data Use Agreement. The dataset is available under the PhysioNet Credentialed Health Data License, which permits academic research use but prohibits re-identification, public redistribution of raw records, or commercial reuse. The data analysis is conducted in accordance with PhysioNet's data-use and security requirements, as well as applicable institutional review and ethical research policies.

The replication scripts and modeling artifacts developed for this study are hosted in a private GitHub repository (https://github.com/NU-Academics/PII-Detection) maintained under the NU-Academics organization. Access may be provided to qualified researchers upon request. The repository is released under an open-source license that applies only to the source code and does not include or distribute the MIMIC-III dataset or any patient-level data extracts

Table 7

Dataset and Code Repository Summary

Category	Description
Dataset Source	PhysioNet — MIMIC-III Clinical Database
Dataset License	*PhysioNet Credentialed Health Data License (research-only use)*
Access Conditions	Credentialed access; completion of data-use training and agreement
Redistribution Status	Raw data may not be redistributed or uploaded to public repositories
Data Protection Restrictions	Re-identification, commercial reuse, and public release are prohibited
Institutional Oversight	Used in alignment with IRB and ethical research guidance
Repository Platform	GitHub — PII Detection Framework
Repository Contents	Replication scripts, modeling pipeline, and configuration files
Code License	MIT License / Apache 2.0 (applies to source code only)
Data Exclusion Statement	The repository contains no original MIMIC-III data and no derived patient-level records.