Grenze International Journal of Engineering and Technology, Jan Issue

A Fully Homomorphic Encryption based approach for Privacy Preserved Pre-processing of MedicalTranscripts

Vijayendra S. Gaikwad[1], Aditya Patil[2], Ruturaj Panditrao[3], Tanisha Pareek[4] and Muskan Agrawal[5]
[1-5]Department of Computer Engineering, SCTR's Pune Institute of Computer Technology, Pune, India
Email: vij711@gmail.com

***Abstract*— Natural Language Processing holds immense potential for extracting insights from healthcare data, but it demands stringent privacy protection. Privacy-preserving NLP techniques, notably Fully Homomorphic Encryption (FHE), offer a path to advanced analytics while preserving patient data confidentiality. This paper marks the culmination of our initial phase, focusing on data preparation and encryption. We've employed FHE with the CKKS scheme to ensure data remains encrypted. Performance evaluation adapts multiclass classification metrics, addressing the distinct nature of healthcare data. As we conclude this phase, we emphasize that this is only the beginning. Subsequent steps will focus on developing NLP models trained on encrypted data. Our work highlights the intersection of data privacy and advanced analytics in healthcare, ultimately benefiting healthcare providers and patients.**

***Index Terms*— Healthcare Data, Privacy-Preserving NLP, FHE, Multiclass Classification, Performance Evaluation.**

I. INTRODUCTION

The advent of Natural Language Processing (NLP) has ushered in a new era of data-driven insights and transformative capabilities across various domains. In the healthcare sector, the promise of NLP is particularly alluring, offering the potentialto unlock valuable knowledge from the vast medical data repositories. However, this promise is intricately intertwined witha profound responsibility—ensuring the privacy and confidentiality of sensitive patient information. Healthcare data, often comprising clinical notes, medical transcriptions, and patient records, is replete with confidential information. The extractionof meaningful insights from this data necessitates advanced NLP techniques. Yet, traditional NLP approaches invariably demanddata exposure in plaintext, posing a significant privacy risk. In response to this dilemma, privacy-preserving NLP techniques have emerged as a pivotal area of research and development. These techniques, notably centered on Fully Homomorphic Encryption (FHE) Ref.[1] Ref.[2], offer a promising avenue to conduct advanced analytics while maintaining the confidentiality of healthcare data. Our endeavor embarks upon this journey with the primary objective of developing a secure NLP system tailored to the unique requirements of healthcare data. This paper presents the culmination of the initial phase of our work. Our focus has been on preparing and encrypting healthcare data, an endeavor that calls for meticulous attention to data quality, confidentiality, and computational efficiency. By applying FHE with the CKKS scheme, we ensure that the data remains encrypted throughout the analytical process, thus mitigating privacy concerns. Furthermore, we delve into the critical aspects of performance evaluation,

Grenze ID: 01.GIJET.10.1.447

recognizing the distinct characteristics of multiclass classification within the healthcare domain. Adapting performance metrics tailored to this context underscores our commitment to rigorous analysis. As we conclude the first phase of our work, we emphasize that this marks only the beginning of our pursuit. Subsequent steps will pivot towards developing NLP models,which will be trained on the encrypted data. The ongoing evolution of our work stands as a testament to the intersection ofdata privacy and advanced analytics in the realm of healthcare. Our fervent aspiration is that this contribution will enhancethe secure utilization of healthcare data, ultimately benefiting healthcare providers and, most importantly, the patients whose welfare remains paramount in our endeavors.

II. Literature Survey

The surveyed paper Ref.[3], delves into the core issue of privacy preservation in Natural Language Processing (NLP), particularly when utilizing deep learning models. It categorizes these models into three groups: data safeguarding, trusted, and verification, addressing the exposure of sensitive information during training and inference. The systematic exploration within these categories reveals methods such as debiasing word embeddings, encryption for secure data handling, and techniques to detect and assess privacy threats, providing a structured framework for addressing privacy concerns in deep learning-based NLP tasks.

Ref.[4] addresses the challenge of privacy in machine learning model training on distributed client devices, emphasizing the growing importance in Natural Language Processing (NLP) applications. The proposed framework employs Randomized Response-Based Local Differential Privacy and Hash-Based Text Representation to enable user participation without compromising sensitive information. This approach not only ensures privacy in Federated Learning but also offers a practical and efficient solution for handling textual data, making it adaptable to various machine-learning tasks beyond NLP. Key advantages include its feasibility, adaptability, and potential transformation of machine learning practices in scenarios where data privacy is critical. However, challenges may arise in implementing and optimizing the proposed techniques for diverse applications.

Ref.[5] provides a thorough exploration of deep learning in the context of natural language processing (NLP) and indirectly addresses privacy concerns associated with deep neural networks. Although the paper doesn't propose explicit privacy solutions, it highlights the risks of deep learning models unintentionally memorizing sensitive data during training, posing a significant threat to user privacy in NLP. The key insight is the necessity to carefully balance model complexity, performance, and privacy concerns. While deep learning enhances NLP capabilities, mitigating privacy risks requires strategic approaches such as model regularization, data preprocessing, and controlled exposure to sensitive information. Achieving this balance is crucial for harnessing deep learning's power while safeguarding user privacy in NLP.

Ref.[6] addresses privacy preservation in natural language processing (NLP) by introducing the Word2Vec model, which tackles the challenge of learning distributed representations of words and phrases. Traditional NLP approaches may inadvertently expose sensitive information in text data, prompting the need for a solution like Word2Vec. This technique transforms words into continuous vectors, preserving semantic relationships while concealing actual content, thereby offering a potent approach to safeguarding user privacy in NLP. The paper also emphasizes the compositional aspect of distributed representations, enabling the encoding of phrases without directly revealing their original content. The core innovation lies in Word2Vec's ability to transform text into a format that maintains meaning while obfuscating sensitive details, forming the basis for developing privacy-preserving NLP techniques.

Ref.[7] introduces the Transformer architecture, emphasizing attention mechanisms' pivotal role in NLP but acknowledging potential privacy concerns due to sensitive content exposure. While not offering explicit solutions, the paper lays the foundation for understanding the complex interplay between attention mechanisms and privacy preservation in NLP. The key takeaway is the importance of balancing privacy and model performance when adapting attention mechanisms to safeguard sensitive information in language processing.

Ref.[8] addresses cancer characteristic extraction using privacy-preserving DL in pathology reports. The central problem is acquiring large labeled datasets for model training, necessitating registry collaboration. The MT-CNN model operates at the pathology report level, ensuring comprehensive analysis of all cancer characteristics. A groundbreaking privacy technique restricts word tokens to prevent PHI inclusion, facilitating model sharing across registries. Key takeaways: collaborative efforts, feasible privacy-preserving distribution, and benefits of novel transfer learning in cancer research and healthcare.

Ref.[9] addresses privacy challenges in Graph Convolutional Networks (GCNs) by proposing a methodology integrating differentially private stochastic gradient descent (SGD-DP) into GCNs. The paper introduces a graph-

splitting algorithm and a differentially private variant of the Adam optimizer, named Adam-DP, demonstrating reduced training epochs and favorable trade-offs between privacy and utility. Experimental results show a 90% relative performance compared to the non-private variant, validating the feasibility of applying DP training to GCNs. This groundbreaking work opens avenues for privacy- preserving applications in various domains, addressing significant concerns in GCNs' utilization in NLP.

Ref.[10] addresses recent data breaches and privacy concerns under GDPR. The central problem is the need for user-centric approaches, prioritizing individuals' interests and compliance. The proposed solution integrates AI mechanisms, including NER, MLP, RF, and Fuzzy Logic, for user-focused data analysis and privacy risk estimation. The "Personal Data Analyser" (PDA), a GDPR-compliant tool, emphasizes user consent and data disposal after each analysis. The solution comprises a hybrid data analysis mechanism and a privacy risk assessment mechanism. Key takeaways include a shift towards user-centric, privacy- preserving solutions, effective real-world applications, and enhanced data privacy prospects.

The problem addressed in Ref.[11] is the efficient evaluation of deep neural models. This work focuses on implementing a standard ResNet-20 model using the RNS-CKKS Fully Homomorphic Encryption (FHE) scheme with bootstrapping. By adopting state-of-the-art approximation methods, the proposed model evaluates non-arithmetic functions, such as ReLU, with sufficient precision. Bootstrapping enables the evaluation of deep learning models on encrypted data. The proposed model achieves a classification accuracy of 90.67% on the CIFAR-10 dataset, which is close to the accuracy of the original ResNet- 20 model with non-encrypted data.

The primary problem addressed in Ref.[12] is architectural complexity. This paper presents a paradigm using Fully Homomorphic Encryption (FHE) that minimizes architectural complexity and removes client-side involvement during the training and prediction lifecycle of machine learning models. The proposed paradigm achieves both model security and data security by employing a no-decryption approach, allowing the server to handle privacy-preserving machine learning (PPML) without rounds of decryption and re-encryption. This paradigm is the first to achieve privacy-preserving decision tree training without decryption while maintaining a simple client-server architecture.

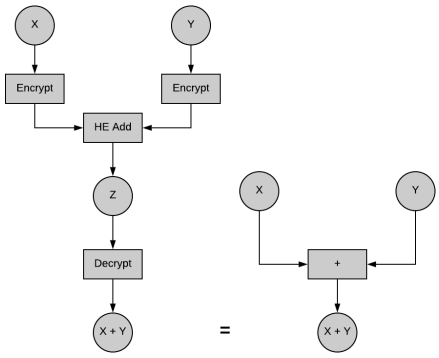


Figure 1: Homomorphic Encryption

III. PRELIMINARIES

A. Privacy Preserving Natural Language Processing

In healthcare, maintaining patient data privacy is paramount. Several techniques are employed in NLP to ensure privacy while extracting valuable insights from textual data. Two key techniques are:

B. Homomorphic Encryption

Homomorphic encryption allows computations on encrypted data, preserving the privacy of sensitive information. The formula for a basic homomorphic encryption function is:

$$E(a) \cdot E(b) = E(a + b) \qquad (1)$$

In NLP, homomorphic encryption can be applied to text data to perform operations like searching and analyzing encrypted patient records without revealing their contents.

1) Homomorphic Encryption Types: There are different levels of homomorphic encryption, each with varying capabilities: Fully Homomorphic Encryption (FHE) Fully Homomorphic Encryption is the most powerful form of

homomorphic encryption. It allows both addition and multi- plication operations on encrypted data. Encryption and encryption are computationally intensive, but FHE enables complex computations on encrypted data. The formulas for FHE operations are as follows:

$$Enc(m1) + Enc(m2) = Enc(m1 + m2) \quad (2)$$

$$Enc(m1) \cdot Enc(m2) = Enc(m1 \cdot m2) \quad (3)$$

Partially Homomorphic Encryption (PHE)

Partially Homomorphic Encryption supports either addition or multiplication on encrypted data but not both. It is less computationally intensive than FHE and is often used in cases where only one type of operation is required. The formulas for PHE operations are as follows:

For addition:

$$Enc(m1) + Enc(m2) = Enc(m1 + m2) \quad (4)$$

For multiplication, if supported:

$$Enc(m1) \cdot Enc(m2) = Enc(m1 \cdot m2) \quad (5)$$

By understanding the capabilities of different homomorphic encryption schemes and applying them in healthcare scenarios, organizations can leverage encrypted data for meaningful insights while upholding the highest data security and privacy standards.

2) CKKS Encryption Scheme: The CKKS (Cheon-Kim-Kim-Song) [13] encryption scheme is a cryptographic method designed to enable approximate arithmetic operations on encrypted data. It supports operations like the addition and multiplication of encrypted messages while also managing the magnitude of plaintext through a rescaling procedure. This rescaling procedure involves truncating ciphertexts into a smaller modulus, which leads to rounding of plaintext. The primary concept behind CKKS is to introduce noise in significant figures of the plaintext, initially added for security purposes, and consider it as a part of the error occurring during approximate computations. This noise is gradually reduced along with plaintext during rescaling. Consequently, the decryption process yields an approximate value of plaintext with a predetermined precision.

3) Key Features of CKKS Encryption Scheme:

1) Batching Technique: CKKS employs a batching technique that is particularly well-suited for RLWE-based constructions. In this technique, plaintext polynomials are elements of a cyclotomic ring of characteristic zero and are mapped to message vectors of complex numbers via a complex canonical embedding map, which is an isometric ring homomorphism.

2) Modulus Growth Control: Unlike some previous encryption schemes, CKKS manages the growth of the bit size of the ciphertext modulus. This growth is linearly related to the depth of the circuit being evaluated, primarily due to the rescaling procedure.

3) Precision Preservation: CKKS ensures that the depth of the circuit bounds the precision loss during evaluation. It exceeds at most one more bit compared to unencrypted approximate arithmetic operations like floating-point operations.

4) Versatile Application: In addition to basic approximate arithmetic operations, CKKS can be applied to the efficient evaluation of transcendental functions such as multiplicative inverse, exponential function, logistic function, and discrete Fourier transform, making it a versatile choice for various computational tasks involving encrypted data.

The CKKS encryption scheme offers a robust solution for performing approximate arithmetic on encrypted data while addressing precision concerns. Its ability to control modulus growth and manage noise through rescaling makes it an efficient choice for secure computations involving approximate calculations.

C. Libraries

1) PySEAL: PySEAL is a powerful library that provides an interface to Microsoft's Simple Encrypted Arithmetic Library (SEAL). SEAL is designed for homomorphic encryption and allows secure computation on encrypted data. PySEAL enables researchers and developers to work with homomorphically encrypted data and perform operations while maintaining data privacy. It offers a wide range of functionalities for secure computations, including the addition and multiplication of encrypted data, key management, and custom operations.

2) TenSEAL: TenSEAL is another library for homomorphic encryption focused on the PyTorch ecosystem. It extends PyTorch to support encrypted computations, making it easier for machine learning practitioners to work with secure data. TenSEAL provides a seamless integration with PyTorch, allowing users to apply homomorphic encryption to PyTorch tensors and models.
3) Concrete-ML: Concrete-ML is a library designed to enhance the transparency and reproducibility of machine learning experiments. It provides tools and utilities for logging experiments, tracking hyperparameters, and managing model versions. Concrete-ML enables researchers and data scientists to document their machine-learning experiments in a structured and organized manner.

IV. METHODOLOGY

The methodology of this paper centers on the implementation of privacy-preserving Natural Language Processing (NLP) techniques in the healthcare sector. Specifically, the paper addresses the secure processing of medical transcriptions, which are vital to the healthcare data landscape. The dataset under consideration is textual, comprising medical transcriptions. The paper's primary objective is to transform the raw dataset into a well-encrypted, vectorized format suitable for NLP tasks.

A. Encryption Technique

To ensure data privacy and security, Fully Homomorphic Encryption (FHE) is chosen as the encryption technique. FHE enables computations on encrypted data without decrypting it, making it a robust choice for preserving the confidentiality of sensitive healthcare information.

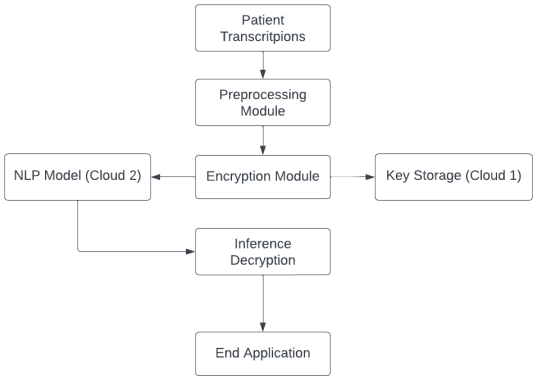


Figure 2: Block Diagram and Data Flow Diagram

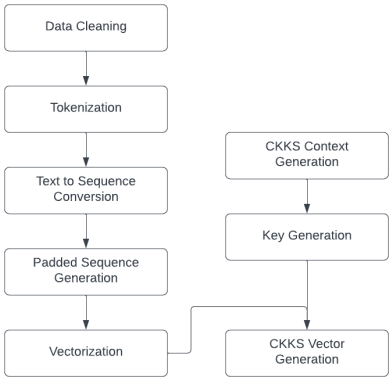


Figure 3: Work Flow Diagram

B. Parallel Study

A parallel study will be conducted to evaluate the chosen methodology's effectiveness. This similar study is inspired by a method outlined in [4], which employed a dataset of IMDB movie reviews. In the cited study, the dataset underwent a series of preprocessing steps, including data cleaning, tokenization, and encryption using a

26-letter-based hash function similar to SHA (implemented with the Gensim library), followed by vectorization using Word2Vec.
The parallel study in our work will provide an alternate methodology using PySEAL, tenSEAL, and Concrete-ML libraries.
1) *Data Cleaning and Preprocessing:* The dataset undergoes a comprehensive cleaning process to remove noise and incon- sistencies, ensuring the quality of the text data.
2) *Tokenization*: Text data is tokenized to break into individual words, making it amenable to further processing.
3*) Sequence Padding*: Well-padded sequences are created to ensure uniform input dimensions. This is a crucial step for many NLP algorithms.
4) *Encryption with tenSEA*L: The dataset, now preprocessed and tokenized, is encrypted using the tenSEAL library, designed for privacy-preserving machine learning in a PyTorch environment.

C. Pseudocode

Algorithm 1 provides the pseudocode, which encompasses the programmatic approach for a full homomorphic encryption using CKKS scheme provided by the tenSEAL library.

Algorithm 1 Pseudocode for Dataset Encryption

```
function SetupTenSEALContext
    context ← CreateCKKSContext
    GenerateGaloisKeys(context)
    SetGlobalScale(context)
end function
df ← ReadCSV('dataset.csv')
CleanDataset(df)
for col in ['description', 'sample name', 'keywords', 'medical specialty'] do
    df[col] ← TransformTextToTFIDFVectors(InitializeTFIDFVectorizer(), df[col])
end for
function ApplyCKKSVector(entry)
    return CreateCKKSVector(context, entry)
end function
df['enc description'] ← ApplyCKKSVectorToColumn(df, 'description')
df['enc sample name'] ← ApplyCKKSVectorToColumn(df, 'sample name')
df['enc keywords'] ← ApplyCKKSVectorToColumn(df, 'keywords')
df['enc medical specialty'] ← ApplyCKKSVectorToColumn(df, 'medical specialty')
Serialize the first entry in the 'enc description' column:
serialized_description ← SerializeCKKSVector(df['enc description'][0])
```

D. Criterion of Evaluation

The primary criterion for evaluating the methodology's success is the computational overhead. Given the healthcare sector's sensitivity and the need for real-time or near-real-time processing, assessing the computational cost associated with privacy- preserving techniques is imperative. Minimizing the overhead while maintaining data security is a critical challenge the paper aims to address.
The proposed methodology is designed to ensure medical transcriptions' privacy, security, and integrity while enabling mean- ingful analysis. It offers a promising approach to leverage advanced privacy-preserving NLP techniques in healthcare data analytics.

V. Empirical Results

A. Dataset

The dataset utilized in this work is the MTSamples dataset, containing medical transcriptions. It consists of 4,999 rows and six columns. The "description" column provides textual descriptions of medical cases, while the "medical specialty" column represents the medical specialty associated with each patient. Additionally, the "sample name" column offers unique names for each sample. The "transcription" column contains the transcriptions of medical cases, although 33 values are missing in this column. The "keywords" column includes keywords related to the medical cases, with 1,068 missing values. Furthermore, the dataset encompasses 40 unique medical specialties,

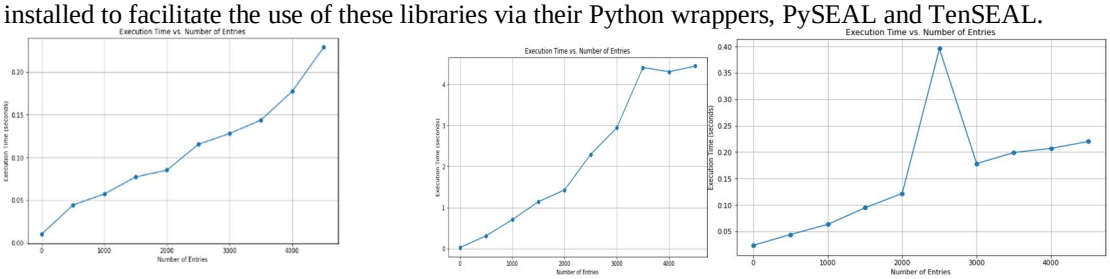
indicating its diversity. It features 2,377 unique sample names and 3,849 unique keywords, highlighting the breadth of medical topics and terms covered within the transcriptions. The MTSamples dataset is a valuable resource for conducting privacy-preserving NLP analyses in the healthcare sector, given its substantial size and comprehensive coverage of medical information. For the coursework of this project, the target variables will be ”medical specialty” and ”sample name,” with other columns serving as features for the analysis.

B. Experimental Setup

This section details the experimental setup used for the work, outlining the environment, libraries, and steps taken to work with privacy-preserving NLP in the healthcare sector.

1) Operating System and Environment: The project was developed on a Linux open-source environment, explicitly using Ubuntu 20.0.4, a stable and widely-used operating system release. This choice provided a reliable foundation for the entire setup.

2) Software and Libraries: To enable privacy-preserving computations and encryption, we installed the necessary software components. This included C++ and CMake, essential for building the SEAL (Simple Encrypted Arithmetic Library) and TenSEAL libraries manually on the local machine. Additionally, Python 3.9 and pip were installed to facilitate the use of these libraries via their Python wrappers, PySEAL and TenSEAL.



(a) Medical Specialty Column (b) Transcriptions Column (c) Keywords Column

Figure 4: Encryption Time Observations

3) Library Testing: Before proceeding with the core project work, we implemented and thoroughly tested the sample code provided by the library documentation. This testing phase ensured the libraries were installed correctly and functioning as expected. It also allowed us to gain familiarity with the libraries and their capabilities.

4) Data Preprocessing: Data cleaning and preprocessing were essential steps in the project. The dataset was meticulously cleaned to eliminate any redundancy or missing values, ensuring the data’s quality and integrity. Imputation techniques were applied where necessary to address missing data points.

5) Setting Up FHE Context: The core of privacy-preserving NLP lies in applying Fully Homomorphic Encryption (FHE). We chose the CKKS (Cheon-Kim-Kim-Song) scheme for our FHE context in this project. This step laid the foundation for secure computations and ensured the data remained encrypted throughout the analysis.

6) Key Generation and Encryption: To enable secure computations, keys were generated within the FHE context using the CKKS encryption scheme with a polynomial modulus degree of 8192 and coefficient modulus bit sizes of [60, 40, 40, 60]. These keys were essential for encrypting and decrypting the data securely. The encryption process involved transforming the data vectors individually, ensuring that sensitive medical information remained confidential.

7) Understanding Encrypted Data: The final step in the setup was to examine how the encrypted data appeared within the FHE context. This understanding was crucial for ensuring that computations and analyses could be conducted while preserving data privacy.

C. Performance Analysis

In our quest to develop a privacy-preserving Natural Language Processing (NLP) solution for the healthcare sector, we must assess the computational overhead incurred at various stages of our work. Computational efficiency is paramount to ensure that our system operates effectively and efficiently while preserving data privacy.

1) *Performance Metrics*: To gauge the computational overhead, we will employ a set of well-defined time-based performance metrics, each tailored to specific stages of our project. These metrics will provide valuable insights into the time-consuming aspects of our system.

2) *Observations:*

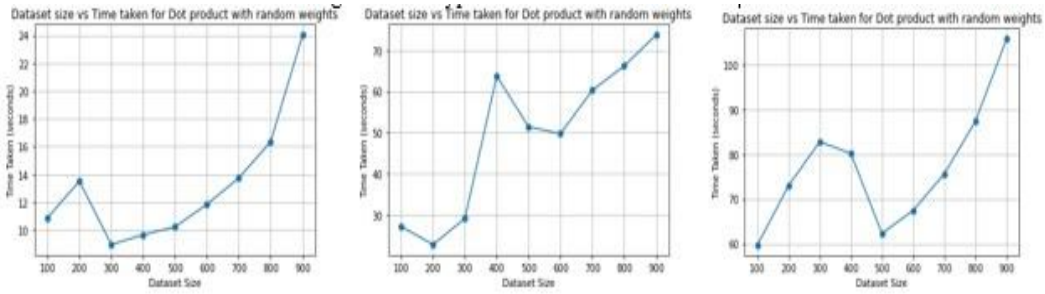
•*Encryption Time*: For the assessment of the effect of the increase in bucket size (dataset), the columns of keywords, transcriptions, and medical specialty were chosen, and the dataset size was increased in steps of

thousands. It was observed that the columns for transcriptions and medical specialty show a closely linear relation, whereas the keywords column gave a sudden spike when the dataset size was changed from 2000 to 3000 entries. Figure4, Figure5, and Figure6 give the visual representations for this experimental result.

•*Encrypted Dot Product Time*: To assess the time taken for the encrypted vectors to undergo a dot product operation with randomly generated weights (values in the range of 1-100), which were encrypted using the same CKKS context. A set of randomly chosen 1000 data points from the dataset was taken, and the size was increased by multiples of 100. It was observed that a similar trend was found in the columns for medical specialty, transcriptions, and keywords. There was an initial linear increase in the time taken followed by a sudden drop, after which the graph nearly resembles an exponential function. Figure7, Figure8, and Figure9 give the visual representations for this experimental result.

VI. Future Scope

In the realm of privacy-preserving NLP within the healthcare sector, our work ventures into uncharted territories with a vision for the future. The upcoming phases will require us to tackle critical challenges. Firstly, we must establish a seamless method for converting encrypted vectors to align with the input stream for our model under development, considering the intricacies of the chosen encryption scheme. Simultaneously, the internal workings of our model, encompassing functions, weight settings, optimizers, and loss functions, will necessitate calibration to harmonize with the encrypted data inputs. Moreover, the storage and management of keys demand thoughtful planning, ensuring model parameters and encryption keys remain securely isolated. Computational overhead is a prime concern, and we are committed to optimizing model training, testing, and validation processes. This will involve exploring techniques like parallel computing and adopting efficient encryption schemes and optimization algorithms. These forthcoming efforts are pivotal to our mission, allowing us to balance data privacy and analytical performance. In doing so, we anticipate delivering a robust and secure privacy-preserving NLP solution that caters to the healthcare sector's unique demands while safeguarding the confidentiality of patient data.



(a) Medical Specialty Column (b) Transcriptions Column (c) Keywords Column

Figure 5: Encrypted Dot Product Observations

VII. Conclusion

In concluding the first phase of our project, dedicated to privacy-preserving Natural Language Processing (NLP) in the healthcare sector, we have successfully prepared and encrypted sensitive medical transcriptions. This marks a significant achievement in ensuring the confidentiality and security of patient data during analysis. The application of Fully Homomorphic Encryption (FHE) with the CKKS scheme has laid a strong foundation for data privacy, enabling secure computations throughout the project. Our experience has underlined the importance of data privacy without compromising the analytical potential of healthcare data. The performance analysis has introduced essential multiclass classification metrics, including precision, recall, accuracy, macro F1, and micro F1. These metrics, adapted explicitly for NLP, offer a holistic view of our system's capabilities. Furthermore, our exploration of computational overhead metrics has unveiled the time investments required at different project stages. These insights will guide optimizing our system for operational viability without compromising data privacy. This first phase serves as a vital stepping stone for the future. Subsequent studies will focus on developing models trained on encrypted data, unlocking the full potential of advanced analytics while maintaining patient data security.

REFERENCES

[1] C. Gentry and D. Boneh, “A fully homomorphic encryption scheme,” in Proceedings of STOC, vol. 9, 2009.

[2] Z. Brakerski and V. Vaikuntanathan, “Lattice-based FHE as secure as PKE,” in Proceedings of CRYPTO, vol. 9, 2014.

[3] S. Sousa and R. Kern, “How to keep text private? A systematic review of deep learning methods for privacy-preserving natural language processing,”Artificial Intelligence Review, vol. 56, pp. 1427-1492, 2023. doi: 10.1007/s10462-022-10204-6

[4] B. Nagy, I. Hegedu˝s, N. Sa´ndor, B. Egedi, H. Mehmood, K. Saravanan, G. Lo´ki, and A´ . Kiss, “Privacy-preserving Federated Learning and its application to natural language processing,” Knowledge-Based Systems, vol. 268, p. 110475, 2023. doi: 10.1016/j.knosys.2023.110475

[5] Y. LeCun, Y. Bengio, and G. Hinton, “Deep Learning,” Nature, vol. 521, no. 7553, pp. 436-444, 2015.

[6] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, “Distributed Representations of Words and Phrases and their Compositionality,” inAdvances in Neural Information Processing Systems 26 (NIPS 2013), pp. 3111-3119, 2013.

[7] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is All you Need,” in Advances in Neural Information Processing Systems 30 (NIPS 2017), pp. 1-12, 2017.

[8] M. Alawad et al., “Privacy-Preserving Deep Learning NLP Models for Cancer Registries,” in IEEE Transactions on Emerging Topics in Computing, vol. 9, no. 3, pp. 1219-1230, July-September 2021. doi: 10.1109/TETC.2020.2983404

[9] T. Igamberdiev and I. Habernal, “Privacy-Preserving Graph Convolutional Networks for Text Classification,” arXiv:2102.09604 [cs.CL], 2021. Available:https://arxiv.org/abs/2102.09604

[10] P. Silva, C. Gonc¸alves, N. Antunes, M. Curado, and B. Walek, “Privacy risk assessment and privacy-preserving data monitoring,” Expert Systems With Applications, vol. 200, p. 116867, 2022. doi: 10.1016/j.eswa.2022.116867

[11] J.-W. Lee, H. Kang, Y. Lee, W. Choi, J. Eom, M. Deryabin, E. Lee, J. Lee, D. Yoo, Y.-S. Kim, and J.-S. No, “Privacy-Preserving Machine Learning

[12] with Fully Homomorphic Encryption for Deep Neural Network,” arXiv:2106.07229 [cs.CR], 2021. https://arxiv.org/abs/2106.07229

[13] H. Lee and C. Clark, “Privacy Preserving Decision Tree Training and Prediction via Fully Homomorphic Encryption with No Decryption,”https://arxiv.org/pdf/2303.01254.pdf