IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

LUNG CANCER PREDICTION USING HOMOMORPHIC ENCRYPTION

Predictive Analysis for Early Detection

[1]Vedika Patil, [2]Aditya Akul, [3]Tanvesh Vaity, [4]Mayur Bharambe

[1]Department of Computer Science & Engineering,
[1]K.C. College of Engineering and Management, Thane, India

Abstract: This study introduces an innovative approach to predict lung cancer using homomorphic encryption and a Convolutional Neural Network (CNN) model Lung cancer stands as a prominent cause of cancer-related fatalities globally, emphasizing the critical need for early detection to enhance patient outcomes. However, leveraging sensitive medical data for predictive modeling poses substantial challenges to privacy and security. Homomorphic encryption (HE) emerges as a promising solution, enabling encrypted data analysis without compromising patient privacy. This report delves into the application of HE in lung cancer prediction, showcasing its potential to revolutionize healthcare analytics while upholding data confidentiality. Homomorphic encryption (HE) serves as a cryptographic technique facilitating computation on encrypted data. In this context, researchers can deploy homomorphically encrypted machine learning (ML) models on public servers without exposing input details, interactions, or model outputs. The initial Fully Homomorphic Encryption (FHE) scheme, proposed by Gentry et al., faced computational overhead challenges for real-world applications such as ML-based inference. During inference, an ML model engages in linear operations, including matrix multiplications and additions between the data matrix, weight matrix, bias matrix, and non-linear operations on the resulting output. Linear operations in HE become more resource-intensive with an increasing number of features, making encrypted computation impractical as the dataset's information content grows. Genomic datasets inherently possess high dimensionality, and the existing privacy-preserving computation literature is limited in its exploration of these datasets. This report explores the potential of HE to address these challenges and open new avenues for secure and privacy-preserving analysis of genomic data in the context of lung cancer prediction.

I. INTRODUCTION

Lung cancer is a prevalent and deadly disease, with early detection being critical for successful treatment. Predictive modeling techniques offer a means to identify individuals at high risk of developing lung cancer, facilitating early intervention and personalized care. However, the use of sensitive medical data raises concerns regarding patient privacy and data security. Homomorphic encryption presents an innovative approach to address these challenges by enabling computation on encrypted data, thereby preserving privacy throughout the predictive modeling process. Homomorphic encryption allows computation on encrypted data without the need for decryption, thus enabling secure data analysis while maintaining confidentiality. By applying homomorphic encryption techniques to lung cancer prediction models, sensitive patient data can be protected throughout the analysis process, from data collection to model training and inference.By leveraging homomorphic encryption, this approach enables healthcare providers and researchers to perform predictive analytics on sensitive patient data without sacrificing privacy or security. Furthermore, it lays the groundwork for developing secure and privacy-preserving predictive models for other types of cancer and medical conditions

II. HOMOMORPHIC ENCRYPTION

Homomorphic encryption is a cryptographic technique that allows computations to be performed on encrypted data without the need for decryption. This means that data can remain encrypted throughout processing, providing a high level of security and privacy. There are various forms of homomorphic encryption, including partially homomorphic encryption (PHE) and fully homomorphic encryption (FHE), each offering different levels of computational capabilities. Utilizing Homomorphic Encryption for Lung Cancer Prediction: In the context of lung cancer prediction, homomorphic encryption offers several advantages:
1)Privacy Preservation: Patient data, including genetic information, imaging scans, and clinical records, can be encrypted before analysis, ensuring confidentiality is maintained throughout the predictive modeling process. 2)Secure Collaboration: Healthcare institutions and researchers can collaborate on predictive modeling projects without directly accessing each other's sensitive data.

Encrypted data can be shared and analyzed securely, fostering collaboration while protecting patient privacy. 3)Scalability: Homomorphic encryption enables scalable predictive modeling by allowing computations to be performed on encrypted data in a distributed manner. This scalability is essential for analyzing large-scale healthcare datasets efficiently. 4)Compliance: By encrypting patient data before analysis, healthcare organizations can ensure compliance with data protection regulations such as HIPAA (Health Insurance Portability and Accountability Act) and GDPR (General Data Protection Regulation).

III. RESEARCH METHODOLOGY

The methodology section outline the plan and method that how the study is conducted. This includes Universe of the study, sample of the study,Data and Sources of Data, study's variables and analytical framework. The detailsare as follows;

3.1 Population and Sample

The study population consists of individuals with diverse medical backgrounds. A sample of patients with documented lung cancer cases is selected for training and testing the predictive model.It Refers to the broader group from which the sample is drawn, potentially representing all individuals within a certain demographic or those with specific characteristics related to lung cancer.A subset of the population that is studied to develop and test the lung cancer prediction model. This could include patients diagnosed with lung cancer or those at risk, with data processed using homomorphic encryption to ensure privacy and security.

3.2 Data and Sources of Data

The collected medical data undergoes a meticulous preprocessing stage to ensure its suitability for analysis. This involves addressing missing values, normalizing variables, and performing feature selection. Any inconsistencies or outliers within the dataset are identified and appropriately managed to enhance the robustness of subsequent analyses. The preprocessing step aims to create a refined and standardized dataset conducive to accurate lung cancer prediction.

3.3 Theoretical framework

The research employs a CNN model for its ability to analyze complex patterns in medical imaging data. Homomorphic encryption secures patient data during both training and prediction phases. The study focuses on variables such as nodule size, density, and other relevant features for accurate predictions.

3.3.1 MACHINE LEARNING ALGORITHMS

1. Convolutional Neural Network(CNN):

The Convolutional Neural Network (CNN) employed for lung cancer prediction is designed with a specific architecture tailored to the intricacies of medical data analysis. The model comprises convolutional layers responsible for feature extraction, pooling layers for down-sampling, and fully connected layers for classification. The number and type of these layers are carefully chosen to optimize the model's performance. Activation functions, pivotal for introducing non-linearities, are strategically placed to enhance the model's capacity to capture complex patterns within the data.
Privacy-preserving cancer inference. Model selection for cancer prediction. While our encodingscheme and feature selection methodologies are targeted towards reducing the number of computations in the encrypted domain, we

follow a similar philosophy to select our ML model. We perform a grid search over several small ML models like Support Vector Machine (SVM) (with radial basis function, polynomial and linear kernels), logistic regression.

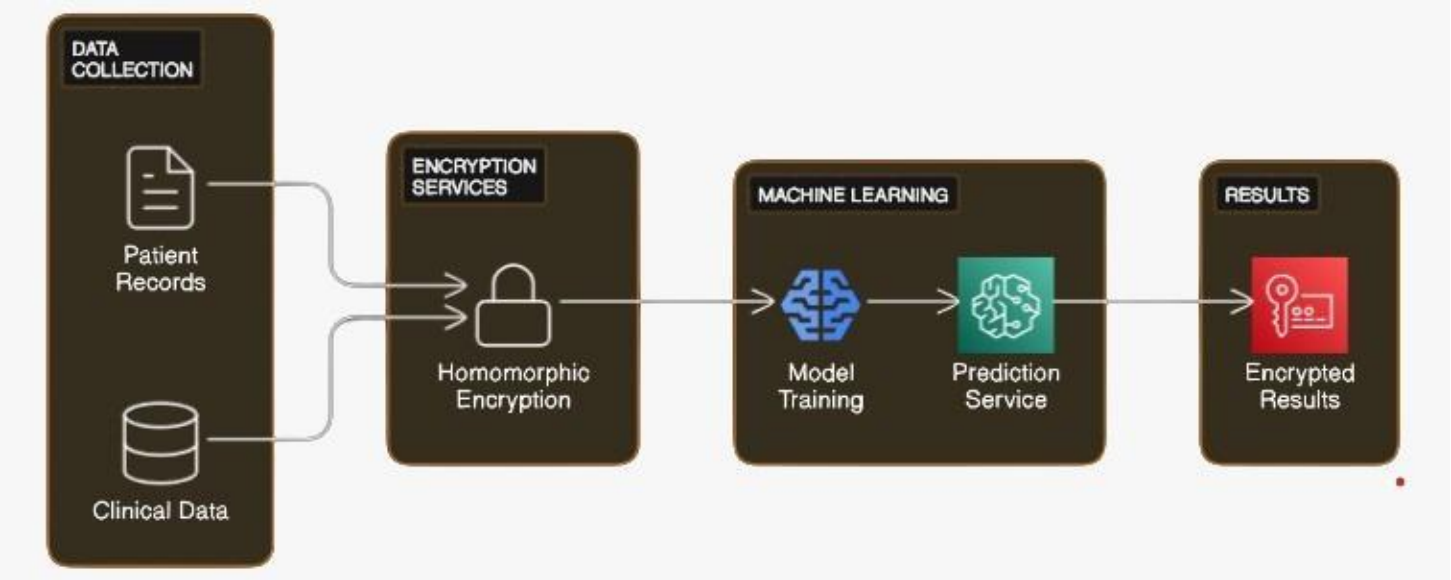

Figure no 2: Model of Homomorphic Encryption

Partial homomorphism started with Rivest, Adleman and Dertouzos. They provided and asymmetric algorithm that supported arbitrarily unlimited multiply operations over encrypted data. This property results into privacy homomorphism as multiply operation can be applied while maintaining privacy of data. Prime factoring problem was the main crust of the algorithm.. Homomorphism in RSA Algorithm Consider two messages m1 and m2 that are encrypted using RSA algorithm with encrypted key as e, where n = p ∗ q, p and q be two large prime numbers. These encrypted messages are multiplied together and proves homomorphic property as shown below: Enc(m1) ∗ Enc(m2) = (me 1(modn)) *(me 2(modn)) = (m1m2) e (modn)= E(m1m2).

Deep Neural Networks (DNNs with two fully connected hidden layers with relu activation) as possible classification models (see Supplementary Information). To capture the non-linearity in the data, we train these models with their respective non-linear activation functions. However, after model selection, we implement the privacy preserving model using the approximation later discussed in “Approximation of nonlinear function in private inference” section. We start with the top 1000 features selected by $\chi 2$ test and increase the number of features by 1000 in each iteration. In our search for the best-performing model, we train models using different numbers of features, different statistical tests for feature selection, different kernels (if applicable), with several regularization techniques, and with different optimization techniques cross-validated over fivefold. Although we train all the models referring to our grid search, we only evaluate and report the best performing models under each category of model and features in the “Results” section.

2. Rivest–Shamir–Adleman(RSA):

In our proposed lung cancer prediction framework, the security of patient data is paramount, and to achieve this, we integrate the RSA encryption algorithm as a fundamental component. RSA, renowned for its robust security features, is employed for encrypting both data and cryptographic keys, reinforcing the protection of sensitive medical information. Seamlessly integrated into the homomorphic encryption implementation, RSA functions to encrypt and secure cryptographic keys, adding an extra layer of safeguarding for data and model parameters. The methodology elucidates the key generation process, ensuring the secure creation of public and private keys and detailing their roles in the overall encryption process. The steps involved in encrypting and decrypting data and model parameters using RSA are outlined, addressing potential computational challenges and proposing strategies for mitigation to enhance overall efficiency.

Metrics to detect problems of unbalanced data :

High test accuracy on unbalanced datasets (with a higher percentage of samples from a particular label) can give a false sense of performance as a random guess (of the label with the highest number of samples) may also result in a high accuracy. For a holistic performance evaluation of our classifiers, we plot Receiver Operating Characteristics (ROC) Curve and report the individual area under curve for each class and the Micro-average Area Under Curve (MAUC) for the classifier. Since ROC curves reflects the entire range of probability threshold, it is a more robust metric and is used in genetic analysis.

Partial homomorphism started with Rivest, Adleman and Dertouzos. They provided and asymmetric algorithm that supported arbitrarily unlimited multiply operations over encrypted data. This property results into privacy homomorphism as multiply operation can be applied while maintaining privacy of data. Prime factoring problem was the main crust of the algorithm.. Homomorphism in RSA Algorithm Consider two messages m1 and m2 that are encrypted using RSA algorithm with encrypted key as e, where n = p ∗ q, p and q be two large prime numbers. These encrypted messages are multiplied together and proves homomorphic property as shown below: Enc(m1) ∗ Enc(m2) = (me 1(modn)) *(me 2(modn)) = (m1m2) e (modn)= E(m1m2).

3. Pillar Cryptosystem for Healthcare Data Encryption:

As a complementary approach to reinforce the security of our lung cancer prediction framework, we introduce the Pillar cryptosystem. Operating in tandem with RSA within the homomorphic encryption framework, this dual-layered strategy aims to enhance the overall security posture, particularly in healthcare data encryption. The Pillar cryptosystem, recognized for its efficacy in healthcare data security, encrypts both data and model parameters, providing unique features such as resistance to specific types of attacks. The integration details within the homomorphic encryption framework are discussed, and a comparative analysis between RSA and the Pillar cryptosystem is conducted. Factors considered include computational efficiency, key management, and resistance to potential attacks. Practical implementation considerations, encompassing ease of integration, computational efficiency, and key management, ensure the establishment of a comprehensive and robust security framework for our lung cancer prediction model, safeguarding patient data throughout the entire analysis process.

3.4 Statistical tools and econometric models

The methodology includes descriptive statistics to understand the distribution of data. Fama-McBeth two-pass regression is employed to estimate risk premiums, and Davidson and MacKinnon equation is used for model comparison. Posterior Odds Ratio is calculated for a formal comparison of non-nested models

IV. ALGORITHM:

1. Initialize parameters and homomorphic encryption keys.

2. Encrypt the input patient data using homomorphic encryption.

3. Perform feature extraction and preprocessing on the encrypted data.

4. Select a machine learning algorithm (e.g., logistic regression, decision tree , RSA) for lung cancer prediction..Train the predictive model using the encrypted training data
   a. Initialize model parameters.
   b. Perform iterative optimization using homomorphic operations: - Compute gradients on encrypted data. - Update model parameters using encrypted gradient descent or other optimization techniques .
5. Evaluate the trained model's performance using encrypted test data:
   a. Make predictions on the encrypted test data.
   b. Decrypt the predictions to obtain the final results.
   c. Compute evaluation metrics (e.g., accuracy, sensitivity, specificity) on the decrypted predictions.
6. Provide the decrypted predictions and evaluation metrics for interpretation by healthcare professionals.

7. Gather feedback and refine the predictive model as needed.
8. Repeat steps 2-8 as necessary for ongoing lung cancer prediction tasks.

v. RESULTS AND CONCLUSION

Homomorphic encryption presents immense potential for transforming healthcare analytics, especially in predictive modeling for diseases such as lung cancer. By facilitating computations on encrypted data, homomorphic encryption empowers healthcare organizations to leverage sensitive medical data while safeguarding patient privacy and ensuring data security. The prediction task focuses on determining whether lung tumors are benign or malignant. While treatment options vary based on tumor type, benign tumors may be managed based on location and symptoms, while malignant tumors necessitate treatment. As advancements in homomorphic encryption continue to progress, its integration into healthcare is anticipated to expand, offering novel opportunities to enhance patient outcomes and drive medical research forward..

```
Test Accuracy of 0: 100% (24/24)
Test Accuracy of 1: 96% (81/84)
Test Accuracy of 2: 91% (103/112)
0.0 0.0
0.0 0.0
0.0 0.0
0.0 0.0
0.0 0.0
0.0 0.0
0.0 0.0

Test Accuracy (Overall): 94% (208/220)
```

Fig No. 2 : Accuracy of Model

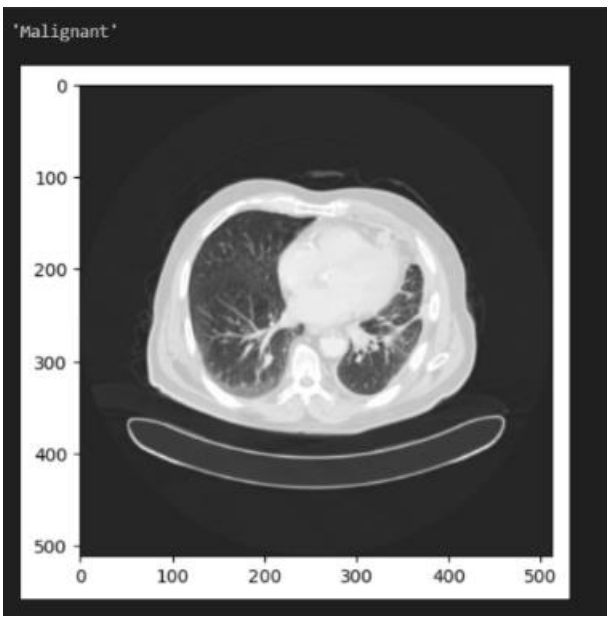

Fig. No. 3 : Predication of Model

VI. ACKNOWLEDGMENT

First and foremost, we extend our deepest appreciation to the patients who generously shared their medical data, making this study possible. Their participation and willingness to contribute to scientific research are truly commendable. We wish to express our heartfelt gratitude to the individuals and institutions who have played a pivotal role in the success of this research endeavor.

VII. REFERENCES

[1] Gentry, Craig. "Fully Homomorphic Encryption Using Ideal Lattices." STOC'09.

[2] Juels, Ari, and Burton S. Kaliski Jr. "Pors: Proofs of retrievability for large files." CCS'07

[3] Ayday, Erman, et al. "Privacy-preserving computation on genomic data: A review of methods." IEEE/ACM Transactions on Computational Biology and Bioinformatics (TCBB) 11.3 (2014): 645-662.

[4] Krishnaiah, V., G. Narsimha, and Dr N. Subhash Chandra. "Diagnosis of lung cancer prediction system using data mining classification techniques." International Journal of Computer Science and Information Technologies 4.1 (2013): 39-45

[5] Zhang, Junjie, et al. "Pulmonary nodule detection in medical images: a survey." Biomedical Signal Processing and Control 43 (2018): 138- 147.

[6] Fenwa, Olusayo D., Funmilola A. Ajala, and A. Adigun. "Classification of cancer of the lungs using SVM and ANN." Int. J. Comput. Technol. 15.1 (2016): 6418-6426.

[7] D chavan,A kiran. “Dot Diffusion Block Truncation Coding For Satellite Image retrieval.”International Journal of Computer

[8] Application 124(4),24-29.

[9] D Chavan ,KA Bhandari. “A New block truncation coding nbtc for satellite image Retrieval using dot diffusion.”USER 6(ISSN2229-5518),