Privacy Aware Deep Learning Model for Multi Class Classification in Big Data

[1]Jaya Sharma and [2]Franklin Vinod D
[1,2]Department of Computer Science and Engineering, Faculty of Engineering and Technology,
SRM Institute of Science and Technology, Delhi-NCR Campus, Ghaziabad, Uttar Pradesh, India.
[1]jayashaa07@gmail.com, [2]datafranklin@gmail.com

Correspondence should be addressed to Franklin Vinod D : datafranklin@gmail.com

Article Info
Journal of Machine and Computing (https://anapub.co.ke/journals/jmc/jmc.html)
Doi: https://doi.org/10.53759/7669/jmc202505073
Received 22 April 2024; Revised from 10 July 2024; Accepted 18 December 2024.
Available online 05 April 2025.

Abstract – Big data and deep learning (DL) are evolving technologies applied extensively in the medical field. Artificial intelligence (AI) technologies have simplified operations such as sharing and retrieving large medical images and swiftly providing disease results in no time. Sharing medical images that are highly sensitive information for every user might give away vulnerable information to the opponents. Privacy is a major concern between a user and a database. In this paper, we propose an Advanced Convolutional Neural Network (ACNN) for selecting the features from large-scale medical data, integrated with privacy-preserved cosine similarity (PPCS), to find similarities between users and all databank images securely. A comparison is made between an ACNN and a PPCS-ACNN based multi-class classification model for diagnosing various lung diseases from Computed Tomography (CT) images. The analysis focuses on the trade-offs between data privacy, diagnostic accuracy, and the efficiency of classification.

Keywords – Privacy Preserving, Image Classification, Big Data, Deep Learning.

I. INTRODUCTION

With the expansion in computational power and production, the amount of data from development in mobile applications, cloud computing technology, social media, shifted businesses in online mode, etc., can rise to big data. So, the data accumulated from diversified sources has resulted in high velocity, high volume and a variety of data, giving birth to a term called "Bigdata". Mining techniques are available to handle data in the developing world, but dealing with big data is still challenging. These are five constituents that represent big data that are high volume (defines the Data capacity), variety (defines types of Data), velocity (describes Data processing speed), veracity (Data trustworthiness and Data legitimacy) and value (usefulness of the Data) [1].

One of the main applications of big data analytics is in Health care applications. The maximum data has been generated in health care through electronic devices [2]. The rapid development of camera technology and the upgrading of digital devices helped produce exponential data growth regarding medical images. In the purview of image classification, certain Imaging tests such as Magnetic resonance imaging (MRI), computed tomography images (CT-Scans), X-rays (electromagnetic waves) and ultrasound are required to be performed. The images of individual organs must be captured to identify the current state of your body's organs and tissues and how they function[3]. In the present climate, almost all hospitals have adopted digital images to diagnose the disease intensity of the patients. Due to the advancement of digital images, image classification has become a significant role with the big data of the medial images. The most appropriate medical images are allocated as same class labels according to their similarities in image classification. It is to some degree seems impossible (in terms of time) to classify these images manually by the doctors because of its large size. In this paper we are using lung CT images to classify lung diseases without compromising user information.

The main aim of medical image classification is accuracy, which is how accurately we classify the CT lung images according to their relevant classes. In the last few years, deep learning techniques have significantly classified structured and unstructured data. Including several deep learning classifiers, the Convolutional Neural Networks (CNN) performs strikingly in medical image analysis activities. Deep learning approaches work on deep architectures to extract relevant features from the image datasets and classify them based on similar classes to diagnose diseases and predictions[4], [5]. In the big data dataset, the images can have several features. That can affect or degrade the performance of our classification model if we give all features into the model. In addition, features can be interconnected, insignificant and redundant, adversely adding noise to the computation time. Feature selection techniques are introduced to overcome the noise problem that can remove insignificant features without affecting the other (relevant) information [6]. The major responsibilities of feature selection are that it can sidestep the curse of dimensionality, trim the training model's runtime, enhance the data's compatibility with the learning model and provide a smoother interpretation of the models [7].

Feature selection approach can work on freely available data stored in a centric database. This approach will also not be successful when data is created and handled by different sources with privacy-sensitive information and do not want to share it. We are sharing or dealing with medical images, which is highly sensitive information for every user and might give vulnerable information to the opponents. For example, when a user interacts with a model by submitting requests for matching features from a database, the model may discover uncommon inside knowledge about the user in question and vice versa. Therefore, privacy and security are more significant features in big data. In this paper, we have constructed a Privacy-preserving framework that identifies similar features without knowing the inside information of the users. We integrate the privacy-based cosine similarity with the ACNN model to achieve this.

The structure of this paper: we discuss the literature review in Section 2, and in Section 3, we propose an integrated algorithm that is privacy-based cosine similarity with ACNN for securely extracting features and delivering effective classification. The explanation of the suggested system in terms of Experiment results is covered in Section 4.

II. LITERATURE REVIEW

The author [8] proposed a full homomorphism encryption approach for extracting features in privacy preservation. They looked at feature selection on distributed datasets as an issue of protecting privacy; imagine that A_1 and A_2, two parties who are only partially truthful, each have personal databases designated DA_1 and DA_2. Addressing the issue of the feature selection for $DA_1 + DA_2$ without jeopardizing their privacy is the objective of the author. The suggested approach may mimic the CWC (Combination of Weakest Components) algorithm on cypher text. The suggested technique reduces computational complexity and resolves the problems with feature selection for a range of original data in a reasonable amount of time. It is among the top performers for the plaintext feature selection problem.

The author [9] proposed a Harmony search technique that uses cosine similarity to get feature selection and facial emotion identification. The author provides the supervised filter harmony search algorithm (SFHSA) for feature selection (FS) based on cosine similarity and minimal-redundancy maximal relevance (mRMR). The Pearson correlation coefficient (PCC) is used to assess the viability of the best feature subsets rather than cosine similarity, which eliminates similar features from feature vectors. The Radboud faces database (RaFD), and the Japanese female facial expression datasets (JAFFE) were used as benchmark Facial emotion recognition datasets for the algorithm's evaluation. Regarding face expression images extracted utilizing five feature descriptors including uLBP(uniform local binary pattern), hvnLBP(horizontal–vertical neighborhood local binary pattern), Gabor filters, HOG(histogram of oriented Gradients), and PHOG(pyramidal HOG); have concentrated on reducing the dimension of the feature sets to achieve higher accuracy.

The author [10] presents a general privacy approach- preserving models to detect similar images. This approach enables hiding the query image and the extracted outcome from the matching server. So, the suggested approach can protect people's privacy in situations where image similarity identification is useful for society but overly intrusive on their privacy. The author's suggested plan consists of three essential steps: Feature extraction- Here, to retain a high level of matching accuracy, both parties turn their images into compressed vector form with a fixed size. Distance computation: This matching phase involves computing the distances between each feature vector held by the server and client query vector. It specifically employed Euclidean distance to find the minimum distance between two vectors. The returned result can be a list of all matched photos whose feature vectors fall inside a certain threshold of being close enough to the specified query. As a result, our scheme's overall complexity is 4(m-1) rounds, where m is secure distance computation.

III. METHODOLOGY

Problem Statement

This paper, we provide a feasible solution to the following problem. Suppose any user wants to send lung images as input to identify patient disease with the pre-trained CNN model. In that case, there can be some possibility to sacrifice additional information for both sides. Therefore, before giving it to the model directly to classify the disease, firstly, we need to find cosine similarities of all the feature vectors of the database along with the query images securely. There are two bodies the user (U), which wants to send private images to identify similar features and the database (D), which contains a collection of images. Our objective is to identify the image in D that is most similar to U without compromising the patient's or model's privacy. Image Detection securely (IDS), which we refer to as our protocol, is defined as

$$IDS \rightarrow val; \tag{1}$$

Where val is the return index value of the most similar image from the database, our method is more general to meet the situation by using Privacy-Preserving Cosine Similarity (PPCS). PPCS can help to extract similar features securely from the database. In the model, if the user wants to give a query image (CT image) as input to diagnose lung-related diseases, first it will go to the PPCS algorithm. PPCS finds features securely using cosine similarity with all database images and returns privacy-preserved image in response. Then, A CNN model classified the privacy-preserved image, and the result returned to the user showed in **Fig 1.**

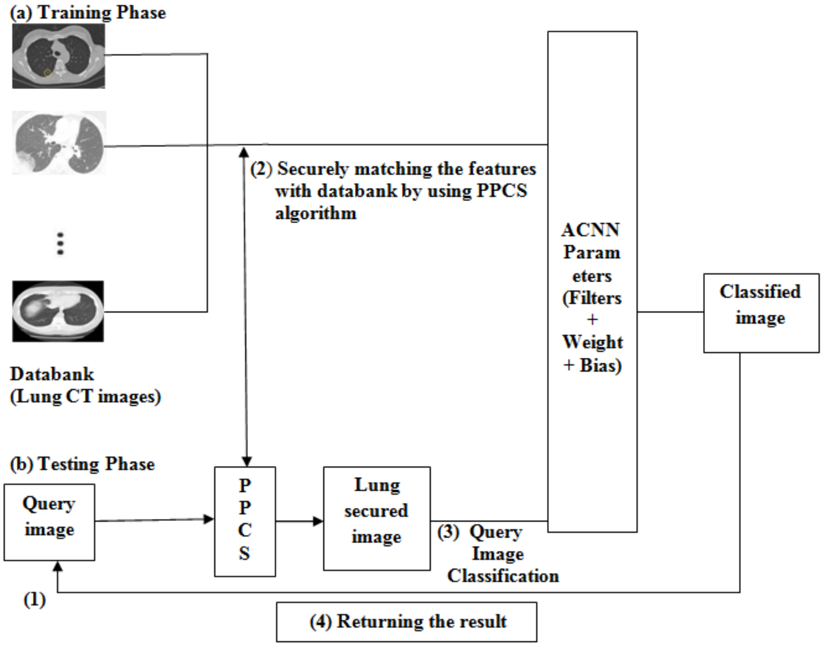
(a) Training Phase
Databank
(Lung CT images)
(2) Securely matching the features with databank by using PPCS algorithm
ACNN Param eters (Filters + Weight + Bias)
Classified image
(b) Testing Phase
Query image
P P C S
Lung secured image
(3) Query Image Classification
(1)
(4) Returning the result

Fig 1. Architecture of Proposed Model.

The subset is created without any original information being sent between the user and the database. It provides enhanced security and protects the users' and the database's privacy while removing any possible security weaknesses.

Proposed Method

Prerequisite Conditions

The following steps are prerequisites to diagnose the disease.

- Data Augmentation
- Training the model (ACNN)

Data Augmentation

A method known as "data augmentation" can increase the number of datasets available for the training in the ACNN model without obtaining new data [11]. DL model training requires a larger dataset, which may be created using a data augmentation approach. Data augmentation techniques include for the model, image resizing, rotation (00 ± 100), scaling (00 ± 200), and shearing (00 ± 100). These methods help to increase the effectiveness of CNNs [12]. The specific specifications of our data augmentation techniques are displayed in **Table 1**.

Table 1. Augmentation Parameters

Augmentation	Parameter
Rotation	± 100
Scaling	±200
Sharing	±100
Horizontal shift	20%
Vertical shift	20%
Horizontal flip	Yes
Vertical flip	Yes

Training the Model (ACNN)

While constructing the predictive model, the feature selection process involves reducing the dimensionality by eliminating irrelevant and redundant features from the input image. In addition, feature selection is the process that selects subclass from the significant features from the input images. In machine learning, some algorithms are available that automatically select features to learn the model. Several other Deep Learning algorithms are often used to train the image model; the ACNN worked efficiently for image-based data processing. An input layer is transformed into an output layer using a series of layers known as ACNN. A group of neurons make up each layer. Each neuron in a layer (apart from the input layer) results from a function $y = f(x)$ applied to the neurons in the layer before it. The fully connected layer (FC), convolutional layer (Conv), activation layer (ReLU), and the pooling layer (Max_Pooling) are a few often employed layers.

Convolutional Layer

The weights and biases shared by the neurons in this layer are frequently referred to as the kernel or filter. If the filter is $\dot{n}$ X $\dot{n}$ in size, an ṅ × ṅ segment of the neurons in the preceding layer will be connected to every neuron in this layer [13]. In a similar fashion, the output for the $(i, j)^{th}$ neuron will be

$$\mathrm{y(i,j)} = \sum_{\mathrm{l=0}}^{\dot{n}-1} \sum_{\mathrm{m=0}}^{\dot{n}-1} w_{l,m} . x_{j+l,j+m} + b \tag{2}$$

Activation Layer

A mathematical function applied to a neuron's output is called an activation function. By adding non-linearity to the model, it enables the network to recognize and depict intricate patterns in the data. Commonly used activation functions include the sigmoid, tanh, and rectified linear unit (ReLU). where ReLU has taken over as the standard recommendation in advanced neural networks. The range of the ReLU activation function is $f(x) = max(0, x)$.

Polling Layer

The previous layer's neurons are divided into a series of non-overlapping rectangles by the pooling layer, and a down-sampling method is used to get the value of one neuron in the current layer from each sub-area. The most typical pooling functions are max-pooling and average pooling. To select output through Max-pooling, select the highest value inside the sub-area. The average polling values for the sub-area will be the output.

The architecture of the Advanced Convolutional neural network (ACNN) is usually a series of Convolutional (Conv)-Activation layer (ReLU) – Pooling layers (Pool) and recites this sequence since images have been converted into a small size, followed by fully connected layer.

The query image has been taken as input for the ACNN architecture. Consider input feature is x', the result of the system for the $(j, k)^{th}$a first hidden neuron is given by Eqn. (2)

$$y' = W'.A + b' \tag{3}$$

Where W' $(0, 1, 2, n - 1)$ is shared weights and bias b. The filter size is ṅ x ṅ, and we use $x_{j,k}$ to indicate the input activation at location j, k. Furthermore, the highest layer of the network is shown in Eqn. (4)

$$Y^l = W^l . A^l - 1 + b^{\,l} \tag{4}$$

$$A^l = g^l(Y^l) \tag{5}$$

Table 2. Layers and Their Parameters of the CNN Model

S.No	Layers	Filter size	Output Size	Parameters/ Dropout Rate
1.	Input	-	224 x 224 x 1	-
2.	Convolutional#1	7x7 /s=2 # 64	112 x 112 x 64	3.41 k
3.	Max_Polling#1.1	3x3 / s#2	56 x 56 x 64	-
4.	Convolutional#2	3x3/ s#1 # 128	56 x 56 x 128	75 k
5.	Max_Polling#1.2	3x3 / s#2	28 x 28 x 128	-
6.	Inception_I1	1x1, 3x3, Max_Pool#1	28 x 28 x 256	263.3 k
7.	Max_Pool #1.3	3x3/ s#2	14 x 14 x 256	-
8.	Inception_I2	1x1, 3x3, Max_Pool#2	14 x14 x 512	77.8 k
9.	Inception_I3	1 x1, 3x3, Max_Pool#3	14 x 14 x 1024	472k
10.	Max_Pooling#1.4	3x3/ s#2	7 x 7 x 1024	-
11.	Dropout	-	7 x 7 x 1024	0.4
12	Fully Connected	-	1 x 1 x 1024	1028 k
13	Soft-max_Activation	Classifier	(None, 3)	-

Then, we flatten the received output, the fully connected layer, to classify our disease. As discussed earlier, this work's primary objective is to guarantee the confidentiality of user lung CT images when employing open DL techniques. Therefore, we suggest using the suggested method to create a special CNN model and evaluate its performance on lung CT images classified as diseases.

Here, **Table 2** depicts all the layers and their parameters used in the ACNN model, that received a 224 x 224 x 1size of the input to identify the diseases. Convolutional layers with 7x7 and 3x3 filter masks and 64 and 128 filters are used, respectively. Followed by inception layers (I1, I2 and I3) are used. In order to reduce dimension utilizing stacked 1x1 convolutions and facilitate more efficient computation in the deeper networks, convolutional neural networks (CNN) use inception modules. The modules are designed to solve a number of issues, such as computational cost and overfitting. After each convolution layer, the max-pooling layers that subsample the images using 2 x 2 filters are added. The activation functions of ReLu are utilized throughout the network. The last two fully connected layers are loaded. After the final pooling layer for regularization, the dropout layer is applied to avoid overfitting. Algorithm 1 follows the classification steps for big data images.

Algorithm 1: To classify big data image

```
Step_1# import pyspark // Session is installed for pyspark
Step_2# import elephas // Enables running large-scale, distributed deep learning models using
        Keras on top of Apache Spark.
Step_3-# from keras import models_Sequential, layers_core_Dense, layers_core_Dropout,layers_core_Activation, optimizers_adam, BatchNormalization.
Step_4# Upload image_dataset into the Spark dataframe and divide it into the training and testing sets // Spark Data_Frame enables the capability to analyze data in various analytical ways.
Step_5# // Layers that added to classifing images
L_1# c1= layers. Conv2D (filters # 64: filter_size#7: s#1: p#same: activation#'relu') (y) // s and p represent as stride and padding respectively.
L_2# c2= layers. Conv2D (filters_f#128: filter_size #3: s#1: p#same: activation='relu')
Layers_ (3,4, &5) # def inception_mode (y: filter_size#1x1: filter_size#3x3: fil_max: name# None)
Result = concatenate ([Conv_filter_size#1x1: Conv_filter_size#3x3: max_pool2D: Dropout]
Layer_Dense // appended_result
Step_6 # Learning and assess the system.
```

Testing Phase

Query Image

Fig 1 shows the basic architecture of the proposed model. In the testing phase, the user gives a CT image as input to diagnose diseases. For further processing, the image is sent to the PPCS algorithm.

PPCS Functionality

The privacy-preserving cosine similarity algorithm calculates similarities between the input image and all existing images in the database without compromising either user personal information or model security. We take the dot product of two images as a vector to find similarities and divide it by the magnitudes of each vector. We compare the input features' cosine values with all the database's image features. The range of the cosine similarity lies from -1 to 1. The angle between two vectors with the same orientation is 0, and their cosine similarity will be high [14].

PPCS Algorithm 2

The following are the steps to the calculation algorithm of the cosine similarity

U_A = User Image A and

$$\text{vector } a = (a_1, a_2, \ldots\ldots\ldots\ldots a_m) \tag{6}$$

D_B = Database Image B and

$$\text{vector } b = (b_1, b_2, \ldots\ldots\ldots\ldots b_m) \tag{7}$$

$$\lambda = lcm(\alpha - 1, \beta - 1), \tag{8}$$

Step1: Key Generation (α, β):

i) Select two distinct prime numbers α, and β and Compute that:

ii) $n = \alpha * \beta$ and Security parameter where lcm operation depict the least common multiple. Let's consider λ is the private key.

iii) Choose a random integer $g \in Z_n^{*2}$

iv) To ensure that n divides g's order, verify that the modular multiplicative inverse listed below.

$$\mu = \mathrm{L}\,(\,\mathrm{g}^{\lambda}\mathrm{mod}\,(\mathrm{n}^2))^{-1}\,\mathrm{mod}\,\mathrm{n} \tag{9}$$

or

We can define the function

$$(\mu)\ =\ {}^{\mu-1}/_{n} \tag{10}$$

(g, n) will be the public key p_K, send $p_K$to U_A for further computation

Step2: Computation on U_A: Encryption (a, p_K)
i) Select a random integer r Є $Z^{*}{}_{n}{}^{2}$(Integer value between 1 and n^2)
for each a_i, i =1, 2..., m
ii)Compute $C_i\ =\ g^a r^n mod\ n^2$
where a = (a_1, $a_{2,}$ ....a_m)
iii)Send $(\mathrm{g}, \mathrm{n}, \mathrm{C_i})$ to $\mathrm{D_B}$

$$\text{Evaluate } A\ = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{a}_{\mathrm{i}}^{2} \tag{11}$$

Step3: D_B Computation (Computed for all database images for input)
For each b_i, i= 1, 2... m

$$D_i\ =\ g^a r^n mod\ n^2 \tag{12}$$

$$B\ = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{b}_{\mathrm{i}}^{2} \text{and} D\ = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{D}_{\mathrm{i}} \tag{13}$$

Send (B, D) to U_A

Step4: Computation by U_A:
Determine $E\ \ =\ r^{-n}.D\ mod\ n$

$$\vec{\mathrm{a}}.\vec{\mathrm{b}} = \textstyle\sum_{\mathrm{i}=1}^{\mathrm{m}} \mathrm{a}_{\mathrm{i}}\mathrm{b}_{\mathrm{i}}\ \ = \frac{E-(E\ mod\ r^2)}{r^2} \tag{14}$$

$$\cos(\underset{\cdot}{\mathrm{a}}, \underset{\cdot}{\mathrm{b}})\ = \cos\frac{\vec{\mathrm{a}}*\vec{\mathrm{b}}}{\sqrt{\mathrm{A}}\sqrt{\mathrm{B}}} \tag{15}$$

The subset is now generated without revealing the model or user details. As a result, there is no chance for a security assault from both directions, and the users and the model's private information are safe.

Query Image Classification

As illustrated in **Fig 2** to create a forward propagation structure, we provide the secured version of the parameters to compute the feature vector of the query picture while maintaining privacy. The feature maps of each layer are thus indicated by the tensor variable v. The convolutional layer and the subsampling layer are in $L-1$ pairs with one fully connected layer. To Utilizing the privacy-preserving primitives for $l\ \le\ l\ \le L-1$, calculate v^l. A trained coefficient is multiplied by v^{L-1}then a training bias is added. Moreover, a sigmoidal function is applied in the fully connected layer. Algorithm1 describes this feature extraction algorithm.

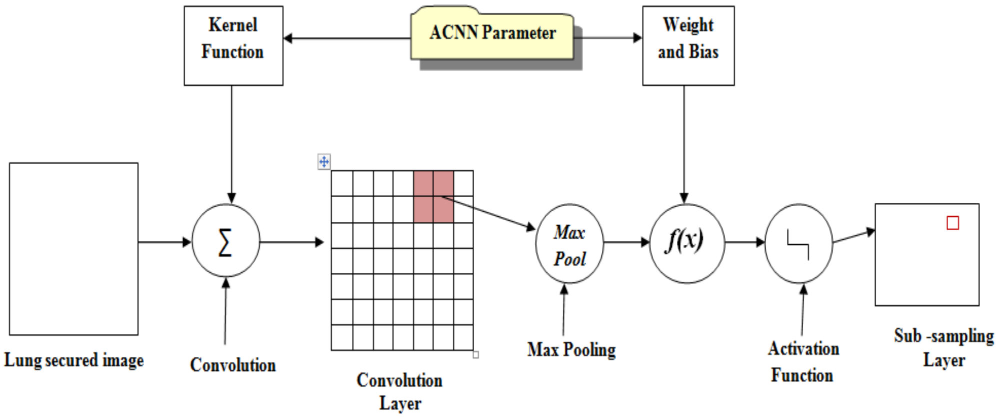

Fig 2. Steps for Extracting the features using ACNN.

IV. EXPERIMENT

In this part, we discussed the Experimental Set-up, Dataset used, as well as how the performance evaluation is conducted for our proposed method.The experimental findings from our proposed work have utilized some of the parameters from [15]. To assess the effectiveness of the suggested technique, various evaluation metrics are utilized in this section to exhibit the performance of the proposed ACNN model with and without PPCS.

The configuration of software and hardware used to implement the proposed scheme are:

Table 3. Illustrate The Hardware and Software Configuration

Operating System	Ubuntu Server 16.04
Memory	8 Gigaoctets
Processor	Intel core i7
Graphics processing unit	NVIDIA
Apache Spark	Spark 2.4
TensorFlow	TensorFlow 2.10.1
Types of Nodes	Master/slave node

Dataset Description

As discussed earlier, it is useful to classify the various disease kinds and look for pertinent examples to use as references for diagnosing diseases using radiological imaging in modern medicine. The primary goal for medical images (health care) is to determine the classification's correctness and assess the search process's effectiveness.

The Following Medical Datasets Have Been Used in Our Evaluation

The Covid-19 Dataset consists of 50,606 labelled lung CT images from several countries, categorized by covid +ve case, covid –ve cases and normal. The dataset collected from different countries is[16], which is from China and includes around 20,000 images in total, [17], Brazil[18], and Iran (Publicly available). **Table 3** Shows Illustrate the Hardware and Software Configuration On the basis of patient level, the dataset was formally divided into a training subset 70% and a testing subset 30%.

For the lung cancer dataset, we used the Lung Imaging Database Consortium (LIDC) and Image Database Resource Initiative (IDRI) [19], which have been widely employed in several research studies [19].Three thousand two hundred fifty lung nodules and 300 CT cases are chosen from these images for the Lung-Deep system evaluation.

Experimental Results

During the search features from the database, the PPCS approach is used. The model determines class based on angles (cosine similarities) to the query image without compromising user information. This approach provides a better way to filter out the unrelated (private) features from the querying image before giving it to the model [20]. Therefore, one of the measures by which our protocol may be evaluated is the performance of categorization. Here, the proposedACNN is the baseline for classification and the performance is evaluated with and withoutPPCS approach.

Table 4. Results Of Different Rounds in The Classification of Lung CT Images

		ACNN			PPCS-ACNN		
		CP	LC	Normal	CP	LC	Normal
20 rounds	CP	**2857**	547	96	**2747**	689	64
	LC	752	**4365**	283	868	**4215**	317
	Normal	508	526	**3966**	665	425	**3910**
50 rounds	CP	**5926**	410	164	**5626**	595	279
	LC	708	**10615**	177	925	**10335**	240
	Normal	665	269	**9566**	823	288	**9389**
100 rounds	CP	**14098**	565	337	**13565**	1065	370
	LC	1465	**22665**	870	1986	**22465**	549
	Normal	1689	980	**21331**	1889	925	**21186**

Table 4 illustrates the classification findings for the medical datasets. Here CP, LC and Normal defined as Covid patients, Lung Cancer and Normal patient respectively. After the model training process, several classification rounds are performed using various test sets of the same size. The PPCS approach reacted similarly to the baseline performance for the original images, as we have shown. It showed that our PPCS methodology can accurately and automatically identify lung diseases (Covid, Lung Cancer and Normal).

As discussed before, the PPCS algorithm securely searches features of the query image with the database. The homomorphic operation affects the time cost of searching features with the database, namely, addition and multiplication. Assuming that a feature vector's length is n. Each cosine calculation includes 4n plaintext multiplication operations and 8n homomorphic addition operations. We used n = 1024 in our work. Therefore, Secure searching time is dependent on n, classes (c), and the total number of images (p) in the database.

Fig 3 demonstrates the duration of secure searching for various c and p parameters. The symbol PPCS(x) indicates that there are x images. We found that when the values of p and c are close together, the searching time reduces. The proposed model estimates the angle of feature vectors, significantly decreasing the search time and improving our approach's efficiency. **Table 5** demonstrates the time required to search for a disease using PPCS and without PPCS.

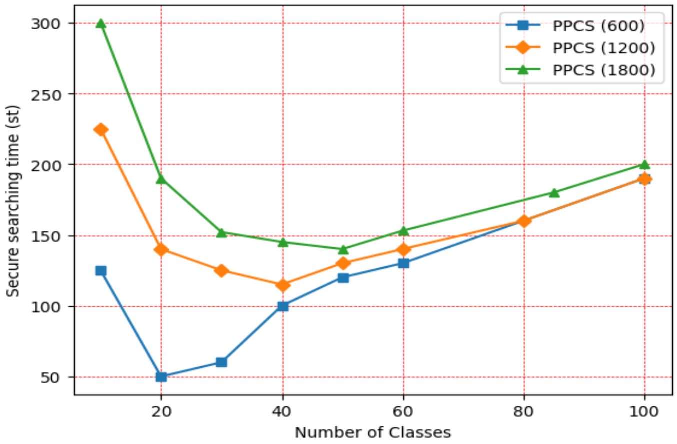

Fig 3. Secure Image Search with Fixed Set of Images.

Table 5. Secure Searching Time of Different Datasets

Dataset	Running time (10^{-3}ms)	
	ACNN	PPCS-ACNN
Covid Dataset	2.3	1.9
Lung Dataset	1.8	1.5
Covid+Lung Dataset	2.6	2.4

V. CONCLUSION

The importance of protecting the privacy of big data image sets is increasing as our healthcare systems become more digitalized and data sharing among healthcare providers becomes more widespread. In this research, we introduced an image-searching strategy with privacy in the big data environment. The semantic security of homomorphic encryption and supervised learning are combined in this protocol to allow for fast and precise image searches in feature maps without compromising the privacy of encrypted data. The experimental findings demonstrate that PPCS-ACNN outperforms previous systems regarding searching time (more than six times quicker) and accuracy rate with privacy on real-world datasets and the ACNN structure.

CRediT Author Statement

The authors confirm contribution to the paper as follows:

Conceptualization: Jaya Sharma and Franklin Vinod D; **Methodology:** Franklin Vinod D; **Software:** Jaya Sharma; **Data Curation:** Jaya Sharma and Franklin Vinod D; **Writing- Original Draft Preparation:** Jaya Sharma and Franklin Vinod D; **Visualization:** Jaya Sharma; **Investigation:** Jaya Sharma and Franklin Vinod D; **Supervision:** Franklin Vinod D; **Validation:** Jaya Sharma; **Writing- Reviewing and Editing:** Jaya Sharma and Franklin Vinod D; All authors reviewed the results and approved the final version of the manuscript.

Data Availability

No data was used to support this study.

Conflicts of Interests

The author(s) declare(s) that they have no conflicts of interest.

Funding

No funding agency is associated with this research.

Competing Interests

There are no competing interests

References

[1]. U. Sivarajah, M. M. Kamal, Z. Irani, and V. Weerakkody, “Critical analysis of Big Data challenges and analytical methods,” Journal of Business Research, vol. 70, pp. 263–286, Jan. 2017, doi: 10.1016/j.jbusres.2016.08.001.

[2]. T. B. Murdoch and A. S. Detsky, “The Inevitable Application of Big Data to Health Care,” JAMA, vol. 309, no. 13, p. 1351, Apr. 2013, doi: 10.1001/jama.2013.393.

[3]. R. Ashraf et al., “Deep Convolution Neural Network for Big Data Medical Image Classification,” IEEE Access, vol. 8, pp. 105659–105670, 2020, doi: 10.1109/access.2020.2998808.

[4]. L. Hall, D. Goldgof, R. Paul, and G. M. Goldgof, “Finding COVID-19 from Chest X-rays using Deep Learning on a Small Dataset,” Apr. 2020, doi: 10.36227/techrxiv. 12083964.v2.

[5]. J. Liu et al., “Applications of deep learning to MRI images: A survey,” Big Data Mining and Analytics, vol. 1, no. 1, pp. 1–18, Mar. 2018, doi: 10.26599/bdma.2018.9020001.

[6]. J. Tang, S. Alelyani, and H. Liu, “Feature selection for classification: A review,” Data Classif. Algorithms Appl., p. 37, 2014.

[7]. Guyon and A. Elisseeff, “An introduction to variable and feature selection,” J. Mach. Learn. Res., vol. 3, no. Mar, pp. 1157–1182, 2003.

[8]. S. Ono, J. Takata, M. Kataoka, T. I, K. Shin, and H. Sakamoto, “Privacy-Preserving Feature Selection with Fully Homomorphic Encryption,” Algorithms, vol. 15, no. 7, p. 229, Jun. 2022, doi: 10.3390/a15070229.

[9]. S. Saha et al., “Feature Selection for Facial Emotion Recognition Using Cosine Similarity-Based Harmony Search Algorithm,” Applied Sciences, vol. 10, no. 8, p. 2816, Apr. 2020, doi: 10.3390/app10082816.

[10]. M. Ravi Prasad and N. Thillaiarasu, “Multichannel EfficientNet B7 with attention mechanism using multimodal biometric- based authentication for ATM transaction,” Multiagent and Grid Systems, vol. 20, no. 2, pp. 89–108, Aug. 2024, doi: 10.3233/mgs-230118.

[11]. H. Phung and E. J. Rhee, “A High-Accuracy Model Average Ensemble of Convolutional Neural Networks for Classification of Cloud Image Patches on Small Datasets,” Applied Sciences, vol. 9, no. 21, p. 4500, Oct. 2019, doi: 10.3390/app9214500.

[12]. S. Hosseinzadeh Kassani and P. Hosseinzadeh Kassani, “A comparative study of deep learning architectures on melanoma detection,” Tissue and Cell, vol. 58, pp. 76–83, Jun. 2019, doi: 10.1016/j.tice.2019.04.009.

[13]. K. Huang, X. Liu, S. Fu, D. Guo, and M. Xu, “A Lightweight Privacy-Preserving CNN Feature Extraction Framework for Mobile Sensing,” IEEE Transactions on Dependable and Secure Computing, pp. 1–1, 2020, doi: 10.1109/tdsc.2019.2913362.

[14]. D. Franklin Vinod and V. Vasudevan, “PPCS-MMDML: Integrated Privacy-Based Approach for Big Data Heterogeneous Image Set Classification,” in Information and Communication Technology for Intelligent Systems: Proceedings of ICTIS 2018, Volume 1, Springer, 2019, pp. 435–443.

[15]. C. Guo, J. Jia, K.-K. R. Choo, and Y. Jie, “Privacy-preserving image search (PPIS): Secure classification and searching using convolutional neural network over large-scale encrypted medical images,” Computers &amp; Security, vol. 99, p. 102021, Dec. 2020, doi: 10.1016/j.cose.2020.102021.

[16]. S. M and N. Thillaiarasu, “A Survey on Different Computer Vision Based Human Activity Recognition for Surveillance Applications,” 2022 6th International Conference on Computing Methodologies and Communication (ICCMC), pp. 1372–1376, Mar. 2022, doi: 10.1109/iccmc53470.2022.9753931.

[17]. “Chest CT Scans with COVID-19.” Accessed: Dec. 22, 2023. [Online]. Available: https://www.kaggle.com/datasets/soham1024/chest-ct-scans-with-covid19

[18]. E. Soares, P. Angelov, S. Biaso, M. H. Froes, and D. K. Abe, “SARS-CoV-2 CT-scan dataset: A large dataset of real patients CT scans for SARS-CoV-2 identification,” MedRxiv, p. 2020.04. 24.20078584, 2020.

[19]. S. Armato et al., “WE-B-201B-02: The Lung Image Database Consortium (LIDC) and Image Database Resource Initiative (IDRI): A Completed Public Database of CT Scans for Lung Nodule Analysis,” Medical Physics, vol. 37, no. 6Part6, pp. 3416–3417, Jun. 2010, doi: 10.1118/1.3469350.

[20]. W. Ning et al., “iCTCF: an integrative resource of chest computed tomography images and clinical features of patients with COVID-19 pneumonia,” Apr. 2020, doi: 10.21203/rs.3.rs-21834/v1.