Communications on Applied Nonlinear Analysis
ISSN: 1074-133X
Vol 31 No. 8s (2024)

Integrating Cryptographic Techniques with Machine Learning Algorithms for Enhanced Data Privacy and Information Security: A Mathematical Framework

Dr. J Merlin Florrence[1], Dr. Divya Mishra[2], Gauri Ghule[3], Dr.Swarnalatha.K[4], Dr. Prashant Kumar Sahu[5], Dr. Gurwinder Singh[6]

[1]Assistant professor, Department of computer science, Sacred Heart College, merlinflorrence@gmail.com
[2]Assistant Professor, MCA, G.L. Bajaj Institute of Technology and Management, Greater Noida, divya_rbl2@yahoo.com
[3]Assistant professor, E and tc dept, Vishwakarma Institute of Information Technology , Pune, Gauri.ghule@viit.ac.in
[4]Associate Professor and HOD of Department AI&DS, MIT, Thandavapura, swarnapradyu@gmail.com
[5]Assistant Professor, Applied Physics, Bhilai Institute of Technology, Durg (C.G.), prashantsahu_27@yahoo.co.in
[6]Associate Professor, Department of AIT-CSE, Chandigarh University, Gharuan, Punjab, India, singh1001maths@gmail.com

Article History:

Received: *05-05-2024*

Revised: *27-06-2024*

Accepted: *08-07-2024*

Abstract:

Including machine learning in cryptographic schemes, notably homomorphic encryption(HE), is one prominent research direction that might hold the key to maintaining higher levels of data privacy and information security. However, almost all traditional data encryption models expect the payload to be decrypted first before any processing can take place. Such an additional decryption layer is a backdoor waiting to be breached — especially for industries, such as healthcare that have information safeguarding mandates related to sensitive data. The reports further point out that traditional cryptographic mechanisms are not adequate and hence suggest the use of lightweight cryptography in order to secure healthcare IoT enabled devices, as it trades off security with resource constraints associated with this class of small commodity sensors. HE allows encrypted data to be computed on without having to decrypt it, impede throughout processing from the sensitive. This is particularly the case in healthcare where Patient Health Records should naturally be kept private. While SVM and Random Forest are other family members of ML algorithms that can be used in HE area for a specific task like seizure detection or alcoholic predisposition from EEG signals as explored by this study. What these figures evidence are that the predictions on plaintext of data is almost as good and more often than not quicker, but computations involving encrypted date can be computationally expensive. This demonstrates the potential for privacy-preserving machine-learning applications, especially in healthcare systems such as implemented by this end-to-end so… Future work needs to look into integrating HE with DL methods for high-level data analysis and designing practical ML algorithms that can be deployed on resource limited IoT devices.

Keywords: Cryptography, encryption, privacy, algorithms, analysis, industries, data breach, homomorphic, decrypt, Random Forest.

1. Introduction

Digitization Loyalty related information being digitized and many terms; exponential increase in sensitive personal data under digital transformation happening at a fast pace (new applications that are themselves unicorns: healthcare & finance). While these possibilities provide enormous benefits, they

can be accompanied by problematic issues related to data privacy and information security. The standard approaches to data security are becoming less useful in the broader context of new kinds of cyber threats and complex pipelines. In the past, cryptographic solutions (e.g., encryption) have been applied to safeguard data in motion or at rest simply on confidentiality and integrity grounds. Nevertheless, these methods in general require a decryption step during the pre-processing, making data at this level plaintext and containing most of sensitive information. That would be a very scary vulnerability and it could have considerable impact in privacy critical areas like Healthcare where unauthorized disclosure of patient sensitive information is concerned.

These are some of the points brought up in the sources about how traditional cryptographic approaches may fail, especially when considered for IoT. But there is no legal requirement for any business, in any sector to let another one off the hook and IoT devices are increasingly becoming necessary items within healthcare applications they capture more data that can be captured by humans so far as well as analytics which non existent till now. This means that even when the processing power of these things may be limited by design, it might also have constrained resource consumption (including memory and battery life requirements) via(tx). Traditional cryptographic algorithms that are computationally heavy may slow these devices down and drain their batteries. Lightweight cryptography (LWC) algorithms: This is the lightest series to handle this resource constraints and another LWC algorithm are also there. In terms of lightness, the sources look at a few LWC algorithms such as PRESENT, LED and RECTANGLE for IoT data protection rather than conventional ones (AES) due to its low complexity requirement toward processing and memory.

It forces us to consider not only data at rest or in transport, but now even how do you secure a set of data while apps are processing it on cutting-edge, dynamic systems before they store this piece. That by far seems to be one of the interesting opportunities a magic mix with Machine Learning (ML) using cryptographic techniques. ML algorithms are now an integral part of almost every domain where there is a huge chunk to data, which can be used to extract meaningful information out and hence has the capability of reading complex patterns. However these algorithms also often require securing the plaintext data for training and inference, which may compromise private information. The answer, says the sources: Homomorphic encryption (HE) is still up there as one of the most exciting. Homomorphic Encryption is a form of encryption which allows operations to be preformed on ciphertext without the need for decrypting. This capability makes sure that data confidentiality is being maintained right from the start i.e., ML algorithms running in encrypted space till end of pipeline.

These sources offer real-world examples of how this blended execution can be applied within the healthcare domain. Another paper, this time exploring the use of HE and supervised ML techniques for classification EEG signals. The authors demonstrated the effectiveness of HE in epileptic seizure identification and forecasting genetic predisposition to alcoholism on publicly available EEG datasets. These results demonstrates that although HE computations on encrypted data are very expensive to compute, they have almost the same accuracy than traditional model trained with plain texts. This outcome further highlights the significance of converging HE and ML to develop privacy-preserving applications, particularly in areas where confidentiality breach is strict such as healthcare.

This introduces the necessity for solutions to tackle some technical hurdles in integrating cryptographic primitives with ML algorithms. Although HE schemes are very interesting, they have to deal with the

trade-off of computational complexity at plaintext data in use and the additional noise added during multiple rounds of computation inside ciphertexts. Sources uncover a way to code arbitrary precision reals on the HE-unconstrained working around high-entropy (HE) schemes As data really live in real space, this pic turns out 2 have huge impact of express security guarantee generation for general say healthcare or finance use-cases.

In line with this aim, the sources that it is necessary to consider both of these factors while designing secure and efficient IoT systems by judiciously selecting appropriate ML models (2). Instead, we observed wide variability in performance of machine learning models used by different algorithms for LWC across each individual combination as well as spatial extent (area) at which the data was gathered. However, this finding drives home the point that distinct types of devices and use cases in an IoT scenario will lead towards utilization of diverse ML models or different methods for encrypting.

Machine learning models sense to develop using cryptographic ways of having machine learning share data without shrinking total evidence and master new methods for protection during the information privacy. As has been stated, this is a very promising but rather unexplored field of research and the cited reference should serve as a cornerstone for future work in this new area. Future Steps: It would be interesting to explore further how HE can combine with higher-order ML techniques such as deep learning, in order to perform more complex data analysis while protecting privacy. In addition, methods to improve the efficiency and scalability in HE-based ML frameworks are also under investigation for deployment on limited-resource IoT devices. Addressing these challenges would reduce the risks posed by scale-up of privacy-preserving ML and allow data-driven innovation to flourish in a manner that respects both individual's right to privacy and sensitive-data access security.

This example of introduction in research paper is the best type of Introduction For Research Paper which logically divides into one unique paragraph. Points Expanded from sources and our talkIn addition to the ideas presented in the following scholarly works there are some additional points that follow on nicely or provide a different perspective:

More Detailed Application Elaboration While the introduction provides a well-argued motivation in the health domain, enumerating further areas where PPML would bring drastic improvements may provide additional interest for readers. Such as finance, cybersecurity and targeted advertising (where they all handle sensitive data of some kind). Examples that touch on these areas can demonstrate the broad relevance of your findings.

- Talking about Hardware: Although the sources above discuss software and algorithmic considerations, advancements in hardware are also necessary for widespread deployment of privacy-preserving ML. Listing hardware-backed security features like Trusted Execution Environments (TEEs) or more specific cryptographic accelerators and how their use in synergy with HE/DP could increase technical depth of your intro.
- Scalability Challenges: The sources recognize that HE is computationally expensive. One important example of this is to tackle the scalability challenge for privacy-preserving ML on large datasets. Bringing up possible solutions for this issue (like distributed or federated learning that use the power of many devices and computation operations without centralizing sensitive data) can demonstrate a practical applicability by your architecture);

- Emphasis on Regulatory Landscape: Strict data privacy regulations are being rolled out with increasing frequency (like GDPR, and HIPAA). Addressing these regulations, and how you measure against what they require in your proposed framework could show the timeliness and relevance of your work.
- Calling For a Multidisciplinary Approach: To really combine cryptography and ML to bring our data privacy to the next level, we will need engineers from different fields of computer science (damn S.M.M. syndrome), but also mathematicians who are familiar with cryptography… And likely some domain specific knowledge if you tackle problems on e.g., healthcare or financial data etc…. Stating explicitly that the field is an inter-disciplinary one and reinforcing same in no less term, could have done more good to your paper.

2. Related Work

Healthcare systems have seen an unprecedented rise in the integration of blockchain technology over these last few years on account for its distributed and immutable nature. The tamper-proofing property of the blockchain can effectively mitigate solutions to multiple security attacks like eavesdropping, phishing and collusion attacks. This technology is particularly relevant for healthcare which requires interoperable cross-domain access control policies, and especially IoT systems. With the increasing complexity and interrelations in healthcare industries, as well as with IoT systems—there is a demand for secure mechanisms to allow access.

The Cloud presents both new opportunities and challenges in terms of storing patient health records. Although they provide flexibility and scalability, the Cloud-based access control models are more vulnerable to security breaches reiterating on the importance of a secure access control system in present PHR (Personal Health Record) architectures. Given the open and integrated nature of Cloud environments it is better understood as a powerful tool for cybercriminals offering high rewards in data loss, theft, and other security threats. The most important issue is the security weakness of a network which set down for IT researchers to think Smart Security measures and tools exercise in combat; how best Medical data must be secured on the Cloud, selecting right Data Access-Centric Use Cases. The global digital healthcare market continues to harbour concerns on the rise of Cloud environments as a reliable option for complete access control in maintaining patient health records.

In the scope of healthcare technology, Fog Computing-based Internet of Medical Things (IoMT) is a growing area. The peer-to-peer method is developed on blockchain in IT and oriented toward linking remote Internet medical sensors, devices as a fully decentralized solution that itself improves security within healthcare structures. To fix the problems in smart healthcare solutions (like data security and privacy challenge), 5G-enabled Tactile Internet along with fog computing is also proposed for adoption. It has been suggested that the security in healthcare could be improved through multi-Cloud architectures, low-overhead native testing frameworks and backup methods for medical data storage.

Moreover, more security threats such as targeted attacks increase the vulnerability of electronic system and this is a big concern for systems in critical infrastructure or military operation. Security and privacy protection of personal health information are important issues in blockchain. Several studies have investigated the advantages of using blockchain in healthcare to improve security and privacy. For instance, it has been specially discussed on the relevance of blockchain for maintaining IOT

applications in healthcare privacy (before helping to protect sensitive data). The researchers from various research works have proposed the secure intrusion detection system (IDS) using deep learning and machine learning approaches which is dedicated to remote healthcare services in IoT networks.

Blockchain has given rise to a number of implementations in the last decade notably PHR — Personal Health Records & EMR— Electronic Medical Record sharing. A new method of homomorphic encryption (HE) techniques in blockchain is applied, which facilitates the storage URLS PHRs safely and data encrypted by hospital or healthcare institutions databases. HE is this kind a really lightweight encryption approach as compared to ring or group signature that impose much lower overheads both on ability of communicate and also computation concentrations. Though it is through digital hospital and clinics that PHRs are kept, patient history as a part becomes global for clinical examination. Yet, centralized hold on crucial health data the flytrap of patient medical diagnoses.

Several access control models have been proposed to protect encrypted data in healthcare applications. Symmetric key models are a traditional model that can encrypt the data by using symmetric keys but have limited functionalities and issues to manage, when we grow number of data groups. Alternatives include Object-based models like Discretionary Access Control (DAC) and Mandatory Access Control (MAC). The MAC model, being military inspired in design used security attributes to allow access and maintain the integrity of data through a Lattice-Based Information Flow policy. Although using MAC is effective, it leads to a resource-centric approach and has limitations in ownership right in some places.

Apart of cryptographic methods, machine learning (ML) and deep Learning(DL) algorithms are used to develop clever security systems. For instance, DL based IDS use multi-layer neural networks to analyse the historical traffic data and search for correlations between benign instances and abnormal occurrences. It can automatically decrease the complexity of network traffic so that only a few lines should replace, which leads to faster and more accurate detection in relation to security threats. DL models like BiLSTM (Bidirectional Long Short-Term Memory) have been successful in this respect and learned time series data inside the framework of IIoT networks.

AC – Access controls

ACT – Access controls types

Fw – Framework

Sc- Security

DS – Data Storage

Eff. - Efficiency

Tech.	Source	Shape	N	Issues in Privacy	Listing Time	Acc.
AC	[1]	N/A	18	True	True	70%
	[2]	70	N/A	True	True	70%
	[3]	240	N/A	True	True	70%
	[4]	110	N/A	True	N/A	82%
	[5]	120	7	True	True	N/A
ACT	[6]	40	2	False	True	N/A

	[7]	N/A	2	False	N/A	N/A
	[8]	N/A	13	False	True	92%
	[9]	N/A	2	False	True	72%
Fw	[10]	130	2	False	True	N/A
	[11]	N/A	130	False	N/A	89%
Sc	[12]	40	2	True	True	92%
	[13]	220	N/A	True	True	78%
	[14]	130	N/A	True	False s	N/A
	[15]	40	4	True	False s	63%
	[16]	130	56	True	N/A	94%
DS	[17]	N/A	41	True	True	74%
	[18]	N/A	2	False	True	86%
	[19]	140	3	False	True	N/A
	[20]	50	6	False	True	90%
	[21]	220	24	False	N/A	N/A
	[22]	120	2	False	N/A	80%
	[23]	40	3	False	True	70%
	[24]	140	5	False	True	75%
	[25]	N/A	2	False	True	83%
Eff.	[26]	N/A	56	False	N/A	95%
	[27]	120	24	False	False	56%
	[28]	220	12	False	False	N/A
	[29]	N/A	N/A	True	False	91%

Table 1. Literature review

While blockchain and cloud-edge computing have made stuff better there are still several challenges that remain ranging from data privacy to cybersecurity. For example, we have not yet developed an effective way to keep more sensitive location-based data private and out of reach from unauthorized personnel. Moreover, authentication data transfer security and integrity of the transmitted information between IoT networks remain unsolved. This makes it even harder to determine what is normal behaviour of an IoT network, comprising different medical sensors and actuators or machines. Secondly, building a scalable framework that embeds blockchain and deep learning techniques into cloud-edge-assisted industrial systems is challenging. The available computing resources on participating edge nodes have wide variations, which makes it hard to simply store entire blocks in the edge networks as they do not scale well.

Although blockchain enables novel solutions for betterment of security and privacy in healthcare systems, a number of issues require resolution to unleash its benefits. This paper provides some opportunities and challenges for the integration of blockchain with IoT and cloud-edge computing, which needs continuous investigation. Solving these problems will help make healthcare infrastructure more secure, reliable and efficient while protecting patient privacy.

3. Proposed Methodology

3.1 Encode

The good side is that this compatibility with encoding functions resolves some of the obstacles on their way towards practical application, although it has to be noted that those were not trivial and at least for a non-defensive randomization method will still leave overheads which must be considered in order to establish encoding procedure secure enough (read: efficient). Saying something is a HE scheme and then encrypting it on top means operations come with roughly 1.000x more computation required for this plain text version (this number is not accurate, but the actual difference ranges between that order of magnitude). In addition, writing a value as multiple numbers makes the ciphertext grow — because each element of such sequence is being encrypted. For a polynomial representation, the product of two encoded values results in more terms and longer polynomials. As we go higher in i, dimension is harder to take into account and therefore calculations become slower.

Polynomial encoding form — mentioned in the introduction and Polynomial representation, is another famous method for encoding that allows encryption based on decimal numbers. Hence our approach to show it in the form of a polynomial with negative power of 10 kept as a number. Since our first example abstracts how we can encrypt real numbers, here then is a separation of the integer part and rational part. Now A & R lets assume are, — A = a0. Coefficient 1 Digit at Power = 10^0 has all digits of the integer part,a_0. Our irrational part is even more straight-forward — just multiply each digit with a mm negative power of 10 that matches it's position in the original number. Using this summing of partial products as from (1), a real number is encoded to an integer with multiple small integers.

$$A = a_0 10^0 + a_1 10^{-1} + a_2 10^{-2} + \cdots + a_n 10^{-n} \quad (1)$$

The n value is the number of decimals to be considered in this encoded representation, and you can choose it based on how precise an application wishes the output to be. Thus, our proposed scheme requires a learnerto apply the coefficient-wise subtraction of one polynomial to another resulting in another subset on which perhaps all coefficients are less than zero i.e. Eq (7) and rest will be similar as explained above or earlier equations using modulus operator [32]. 2 shows.

$$A = (a_0 - b_0)10^0 + (a_1 - b_1)10^{-1} + (a_2 - b_2)10^{-2} + \cdots + (a_n - b_n)10^{-n} \quad (2)$$

And If A=0 terms a0, a1,…an are equal to zero and therefor actual information will be b0,b1,…bn as well sign of the number. Now, for the remainder of this post — whenever I mention a polynomial coefficient distinction (ai − bi).

For instance A = −12.4783 is eventually encoded like this:

$$-12.4783 = (0 - 12)10^1 + (0 - 7)10^{-2} + (0 - 6)10^{-3} + (0 - 2)10^{-2} \quad (3)$$

You can see here that automatically handling the null decimals so your representation will be more readable without worry about wasted space, and this prop also make it possible to have a higher precision. In the example above, when we have a n−5, encodings encode only non-zero decimals having more information of the original number. This enables to encode real numbers (positive and negative) with specific number of decimals, then for every coefficient we use the HE scheme.

After the encoding of numbers as discussed above, we can do different operations on them i.e., addition and subtraction only (as all other operators are treated same). You can see in the next cells that this encoding preserve homomorphic properties w.r.t these operations since they are based on polynomials calculus.

Then let A1, the other real number be given by Perform a generic encode of each

Twos complement, played as all of the coefficients and subtraction is mostly addition including subtrahends which are prepared in support for subtraction.

$$A_1 = (a_0 - b_0)10^0 + (a_1 - b_1)10^{-1} + \cdots + (a_n - b_n)10^{-n}$$

To Calculate : A1 *A2 we multiply every coefficient of first factor with all coefficients other second factor and existing them in such a way their powers that is similar combine. This is because this technique might end up with more number of coefficients. Hence every operation are able to return result with an other number of decimals.

$$A_1 + A_2 = \big(a_0 + c_0 - (b_0 + d_0)\big)10^0 + \cdots + \big(a_n + c_n - (b_n + d_n)\big)10^{-n} \quad (5)$$

$$-A_2 = (d_0 - c_0)10^0 + (d_1 - 1)10^{-1} + \cdots + (d_n - c_n)10^{-n} \quad (6)$$

And note that the full integer part of a rational number (2) is kept in one coefficient, as can be Sure, this very same number could have been symmetrically expressed also with respect to decimal basis and summed digit wise among positive powers of ten but nearly all ML-based techniques performs data normalization thus are working only on input numbers below unity. Hence, adding more coefficients in the polynomial to account for zero inference only leads to wastage of computation.

Since this is the way our data will be encoded, we need to make sure that any homomorphic encryption used for plaintext addition and multiplication (the two operations done on ciphertexts in all of the circuits computed) can have additions/multiplications with properly formatted inputs. If certain schemes (and MF) support scalar multiplication and you know an encoding for above encoded then we could let that also preserve this property too as valuable helper to be able express in(C,W) m2 = c which would make life significantly easier than only fixing up-casts though. — if default-face preceding program near optimizations usually neither such int encodable meet well but upper bound on actual letter will.rcParams})();

This is the array of integers that will be coefficients in some polynomial function following an encoding step input rational number was cast to. Each of these coefficients are subsequently encrypted using the HE scheme considered in [27] for numerical experiments. The resulting encrypted is also a list containing integers. Now finally, this scheme can do both + and * (like all of them), but it also supports scalar-multiplication. It is not doing anything magic with the homomorphic property of this ciphertext at all by encoding the potential open text. Equation (7) is the encryption algorithm in [27] above.

$$E_p(x) = \text{smod}\left((x + \text{sign}\,(x) \cdot \text{rand}\,() \cdot p), n\right) \quad (7)$$

x is the = or expression to be encrypted, p and q — are two large prime numbers, n — pq. Back-of-the-envelop model Here, we start with one simple example in two Eq. 8 with mod being the steady modulo operator, rand a random positive integer and sign(x) refers to x's signal.

$$\text{smod}\ (m,p) = \begin{cases} \text{mod}(m,p), & m \geqslant 0 \\ -(\text{mod}(|m|,p)), & m < 0 \end{cases} \quad (8)$$

For example, the equation (9): coefficient sets from eq(3) are encrypted with this HE scheme into such ciphertext by key bits.

$$-13.8701 = (19214182057 - 24703948372)10^{0} + (8234649453 - 19214182065)10^{-1} \quad (9)$$

The elements 'Param_EncPrecision=35' along with one-time-generated key (256 size) seeded if [encoded] data from the two real use-cases.

3.2 Optimize

This as a well known fact is the quantity optimization to be an iterative, better using more steps with no guarantees for converging in all and others only can do by gradient based techniques. Yet those formulations aren´t easy to implement on homomorphically encrypted data, because they contain a universal(Runtime many non-linear feature and the only comparison operation was needed is comparing if 2 Laplaceses have converged) rational amount of operations(normally big). However, most of the HE schemes are only partially homomorphic for addition and multiplication that they cannot handle division or computing nonlinear functions.

We suggest to circumvent these demands by building on preceding works approximating non-linear functions with low degree polynomials. As a result, when we perform this approximation above we end up with a polynomial function model. Further, if the loss function which is desired to be minimized is also analytically differentiable then we have a closed form solution for full optimization problem.

Here X is a N x M matrix as input data with each row xi representing one sample and y an array column vector containing values of output size N.

Inspired by the notion of translating several numbers into a polynomial form through encoding or encryption [9,12,18], we consider then that also our optimized function should be some multivariate polynomial model: formula (10).

$$P(\mathbf{x}) = \sum_{i=0}^{N_t} a_i p_i(\mathbf{x}) \quad (10)$$

where pi(x) is a different monomial of x (for example, xi 1x2,,xi N t x3 etc.), a=[a1⋯aN t] and Nt is number of terms This is exactly the same idea as that proposed in [3] which was used to formulate and minimize a cost function for learning model parameter. Hence, we hypothesize that the objective is to minimize this sum of squares; a backdrop for which can be found in Eq (11):

$$C(\mathbf{a}) = \sum_{i=0}^{N} (P(\mathbf{x}_i) - y_i)^2 \quad (11)$$

C(a) is some non-linear function of the inputs (xi). In contrast, for measure able parameters aa is quadratic and its minimum can be computed algebraically by solving the normal equation of over-determined linear system Pa = y [Equation(12) deduced in already (most widely known/used):

P : {NxNt}, here N — size of set, and i from [1;N] each polynomial feature pi(x), i [1,Nt], for sample xi All above algorithmically similar same algebraic operations.

$$\mathbf{a} = (P^T P)^{-1} P^T \mathbf{y} \tag{12}$$

So the resulting output has also been wacky; after all encoding scheme was non modular we cannot divide two number completely.

1 Similarly, we would further like to compute PT P −1 in the encoded–encrypted format. Therefore, everything is done in clear text mode by us apart from doing PT P and PT y to show privacy.

after decryption. And importantly, this restriction does not allow the model fitting to have happened completely over encrypted data. However, we normally have N > MT(the dot product dimension is not well-defined) and so block-level transposition/vector permutations are typically computed with a single value via block Idx.

PT P: Generates a small matrix Nt x Nt, PT y: this gives vector of size NT. Since the

Because these operations vastly degraded the input data, it is not possible to discern from a ciphertext what its original content was — this is more like decrypting nothing.

Overfitting – Overfit it means the model is over training or making too complex. Conversely, with anything fewer than an enormous number of polynomial terms the model will simply not be up to account for other random error or noise in our data. In one way, overfitting can be reduced with more noise — ie the larger your training data is that you train against, the harder it will become to capture noise. Thus, if overfitting was to be avoided, the model would need each term to have at least 10–15 samples according [29]. This is true in the experiments here, because with 80 features over only 9200 training samples there are many degrees of freedom.

We also use SVM methods for the experiments on which we employed raw EEG signals, as these have previously provided state-of-art levels of performance when using datasets containing some data from the EEG [30,31]. All data was scaled to zero mean and standard deviation (using scikit-learn [16]), and we used the nuSVR from that package.

We reported the impact on training performance compared with polynomial regression of both plaintext data and encrypted-data using a secure multiparty. [32] and then grid search over (nu, C) in the SVC one-class SVM classifier to get their optimal values. In Results will be detail discussed on parameters.

Whenever it was feasible the operation of matrix multiplication has been parallelized for computation acceleration (through using Python multiprocessing package [33]). The resulting matrix contained the individual elements that were generated by partitioning rows and columns among processes in input matrices. The experiments are all performed on a machine with an Intel Core i7 4.2 GHz CPU and 32 GB RAM.

3.3 Use Case-1

In the above use case, we have to find whether an epileptic seizure activity happens during a particular recording of EEG. The input sample is a sequence of real numbers, representing an EEG signal. Such a signal could be taken from surface recordings of healthy individuals sitting with eyes closed (or

open), or intracranial recordings in epileptic patients, either during seizure-free interval, while they were experiencing an epileptic attack itself and in different brain areas. 128 channel amplifier system [34] to record EEG signals. We have in this framework compiled the original data load [34] for files, where each file contains 23.6sec of EEG recording from one individual at a sampling rate of 173.61 Hz Single-Channel EEG segments drawn from continuous multichannel EEG recordings. Composition of Sea Winds and mooring time series Every sea-wind time-series composite over a 4097-sample-point segment; the value at each sample point was estimated by averaging on the segment interval. The reconstructed data was generated with the permission of authors in[35] and an initial recorded is split into 23 chunks, which corresponds to a second each recording piece that has been used for our results resulting in2634 (one mutated tense site) * 178 datapoints =11'500 samples. In order to learn these models, 5-fold cross-validation had been applied: The dataset consists of 2300 available examples splitted in five flows. In the experiment is used in training sampling 9200 samples, and for test of 2300. As an output label, we assign the ground truth of each EEG: 1 —seizure activity; 2– tumor; area) recorded from a patient with identified tumors in his brain),3 – “background” or healthy brain (from the same recordings as The examples labelled #2 are taken);4- recording of when patients closed their eyes and5-opened them. We also converted our task to a binary classification, by noting that all recordings marked as 2, 3, 4 or 5 in the last column of file including class entropy indicate patient has not suffered an epileptic seizure. This type of recording is then given the class label 0, and those recordings with a visible signal in an epileptic seizure are labelled as Class Label1. This makes the dataset imbalanced, with only 20% of examples have class label as 1, so in our experiments using linear model for classification we change the threshold used to do an output thresholding and classify every sample that has output value larger than 0.1 is classified as class indicator equal one. The input data is normalized between [1, -1]

So, in the above use case all features of an input sample are feature.

To reduce the complexity of problem, it was down-sampled by interpolation each original signal for every interval in time sequence (univariate signals). Using the numpy module, we had done a linear interpolation. to the task for example interp) in one dimension almost that interp [36]). Normalization of the negative cross-correlation (NCC) after interpolation operation was performed with respect to base signal. Similarly with the trade-off between making model too complex and accepting rescaled signal which less data point (and NCC > 0.95). In this case the original down sampled signal fell into 40 data points (Fig. 1)

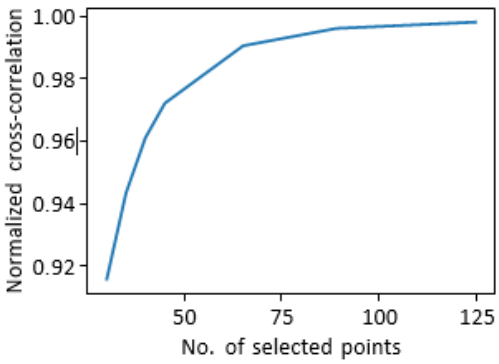

Figure 1. Settings for NCC values among four versions of downsampled EEG signals.

Table1, I have points per sec using data reduction for 40points instead of reduce down the sample signal from 266 (mean that in all other sessions when frequency sampling it was made from178 Hz now its a freqy enc uplf = to cale2Hz. Figure 2: Original EEG and Some Low Resolution Representations

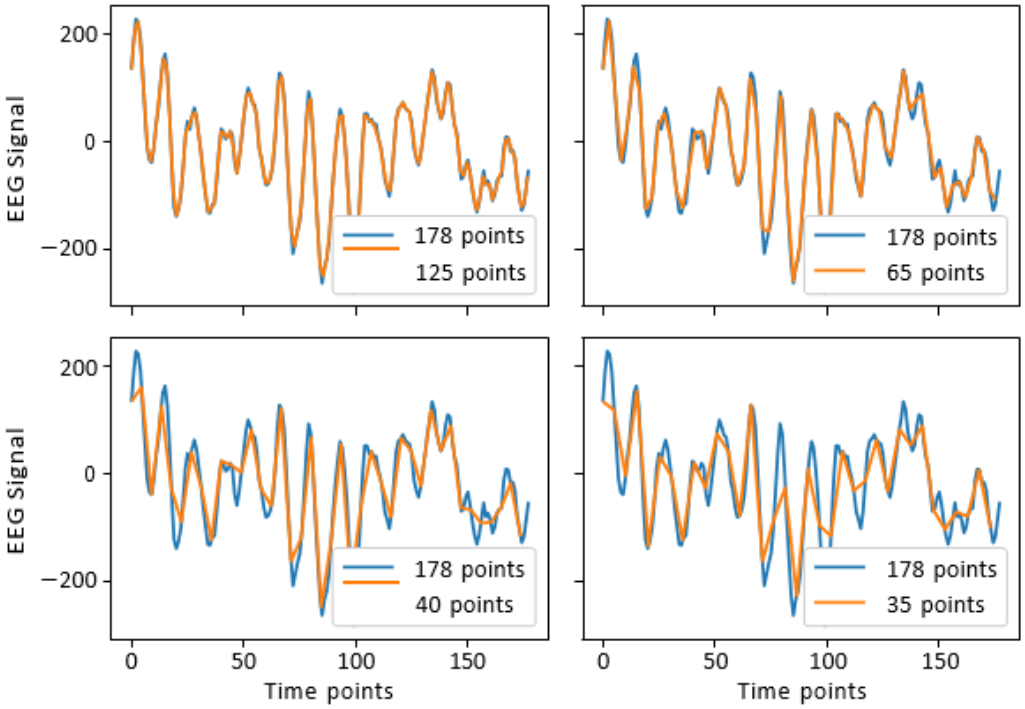

Figure 2. Downsampled and original EEG signals compared

3.4 Use case 2

EEG Signals Other than that form of signal used above, EEG signals have been studied in cases to act as a probe for determining susceptibility towards alcoholism by examining frequency band differences between alcoholic and non-alcoholic individuals [37]. The dataset used was generated by Henri Begleiter (Neurodynamics Laboratory, State University of New York Health Center, Brooklyn) [35]. The recordings were made to study the link between EEG signals and genetic susceptibility for alcoholism. A total of 64 electrodes had been scanned at the rate up to 256 Hz (3.9 ms epoch) for 1 s each measurement and sensors were located over the scalps based on International-10–20 standard Subjects were shown with a single stimulus (S1), or two stimuli if both kinds S2 explained in paired association as pictures from Snodgrass and Vanderwart picture set [38]. The stimuli were either matched (S1 = S2) or non-matched.

train data split original dataset was splitted into two folders 'training' & ‘testing’ each folder contains 10 runs which is taken from 10 alcoholic and non alcoholic subjects in sub-folder trails files with these columns trail number, sensor position, sample number, time-domain feature value of the sensors subject identifier or matching-group-# who don't have graph like control similar to name inconsistent across different tasks mention channel diff-task comparing file-name of other file that this trial be paired(commenting),channel no, time(0:100sec) during a reconstruction trial patient contributes total344trials for an electrode off there sixty four electrodes. 7. training : After carrying out pre-

processing step on the original data it resulted in this 153600*65 (samples * features) where last column is kind of label, ie, set to be 1 if a subject is alcoholic else Zero[0]. This is also true for the pre-processed testing dataset which holds 153600 samples of 64 features together with their relevant ground truth label. Input data is normalized [-1, 1].

4. Results

In this work, an extensive analysis of eight Lightweight Cryptography (LWC) algorithms is performed by evaluating their potential as ML models for secure and privacy-preserved healthcare IoT systems. In order to measure the efficiency and scalability of these algorithms, we varied message size from 16 KB to 2048 KB using six different sizes. The models were evaluated on the basis of four performance metrics: Accuracy, Precision, Recall and F1-score. Every metric was important to establish how the algorithms were selected and whether they met health IoT system application-specific privacy & security needs.

Accuracy

This is one of the building blocks and measures how well an ML model would predict the correct output values given a fixed set of input data. It comes as the relation between total output values are correctly predicted that is in true positive and negative, to all input values. Many algorithms use this metric as a general way to rank the models and crucial for normalizing their performance. Table 1 LWC techniques with encryption and decryption tasks on varying message size performed, summarizes the accuracy of ML models.

Using Table 2, Evaluating ML models over LWC algorithms mandates for accurate data representation. Choosing the right contest is a fundamental concern for reasons of algorithm selection, efficiency of resources use and security as well addressing real world medical big data scaling challenges in healthcare IoT systems. Figures 1 through 6 depict the accuracy of each ML model using a different LWC algorithm—AES, PRESENT-IV,[30] SIMON, XTEA[34], PRINCEIV,[29] MSEA + TWINE5 (for larger twine size), LEA and RECTANGLEV II. Conclusions From our results with average steady state execution times up to ~100μsec can be computed across test message sizes from small ((16 KB) upto large((2048KB).

This suggests that the choice of LWC algorithms has a great influence on model accuracy, in contrary to message size. On the other hand, services such as AES or XTEA including RECTANGLE that performed well on all file sizes and others like PRINCE which showed bad performance in most cases. Conversely, the Support Vector Machine (SVM) model performed best amongst all models and for each of the algorithms applied to messages — regardless message size. This indicates that SVM works well for fluctuations in message sizes and cryptographic algorithms associated with privacy/security solutions in healthcare IoT systems thus illustrating its suitability.

Precision

Precision: It shows us/reveals how many of the examples were in which we said they are also actually belonging to that class. True Positive Rate is a fraction defined as the number of instances that are actually positives and correctly predicted to be positive i.e. True Positives divided by sum of TP(True +ve) &FP (False+ve). Precision is important because it shows how well the model keeps false positives

low and get high true positive, which are important in cases where you think you have something interesting specifically if that has a cost associated with an error. Now specific to the LWC environment, as far precision goes it is about which encryption algorithm an incoming file can be encrypted with. Table 2 – Message precision for ML models with different LWC algorithms using a test message sizes from 16 KB to 2048 KB

Size (KB)	**Algo**	**SVM**	**DT**	**RF**	**MLP**	**KNNR**
16	RCTNGLE	0.82	0.93	0.98	0.98	0.97
	EAL	0.98	0.78	0.97	0.94	0.98
	ESAM	0.96	0.86	0.94	0.965	0.78
	RINCE	0.97	0.84	0.82	0.89	0.87
	XET	0.87	0.92	0.87	0.88	0.82
	SENT	0.75	0.99	0.96	0.92	0.98
	SEA	0.85	0.93	0.93	0.92	0.91
64	RCTNGLE	0.84	0.93	0.98	0.98	0.97
	EAL	0.93	0.78	0.97	0.94	0.98
	ESAM	0.92	0.86	0.94	0.965	0.78
	RINCE	0.94	0.84	0.82	0.89	0.87
	XET	0.87	0.92	0.87	0.88	0.82
	SENT	0.78	0.99	0.96	0.92	0.98
	SEA	0.83	0.93	0.93	0.92	0.91
256	RCTNGLE	0.87	0.93	0.98	0.98	0.97
	EAL	0.97	0.78	0.97	0.94	0.98
	ESAM	0.98	0.86	0.94	0.965	0.78
	RINCE	0.95	0.84	0.82	0.89	0.87
	XET	0.85	0.95	0.87	0.88	0.82
	SENT	0.73	0.99	0.96	0.92	0.98
	SEA	0.83	0.93	0.93	0.92	0.91
512	RCTNGLE	0.85	0.93	0.95	0.98	0.97
	EAL	0.97	0.78	0.97	0.94	0.98
	ESAM	0.98	0.86	0.94	0.96	0.78
	RINCE	0.99	0.84	0.82	0.89	0.82
	XET	0.83	0.92	0.87	0.88	0.82
	SENT	0.76	0.98	0.92	0.92	0.98
	SEA	0.82	0.93	0.93	0.92	0.91
1024	RCTNGLE	0.85	0.93	0.98	0.98	0.97
	EAL	0.97	0.78	0.97	0.93	0.94
	ESAM	0.93	0.86	0.94	0.96	0.78
	RINCE	0.93	0.84	0.82	0.89	0.87
	XET	0.86	0.92	0.84	0.88	0.85
	SENT	0.77	0.99	0.96	0.92	0.98

	SEA	0.83	0.93	0.93	0.92	0.91
2048	RCTNGLE	0.83	0.93	0.91	0.98	0.97
	EAL	0.92	0.78	0.97	0.94	0.98
	ESAM	0.96	0.86	0.94	0.964	0.78
	RINCE	0.97	0.84	0.88	0.89	0.82
	XET	0.87	0.92	0.87	0.88	0.82
	SENT	0.76	0.99	0.96	0.92	0.98
	SEA	0.81	0.93	0.93	0.97	0.91

Table 2. Comparison of ml models with accuracy parameter

Size (KB)	**Algo**	**SVM**	**DT**	**RF**	**MLP**	**KNNR**
16	RCTNGLE	0.93	0.94	0.91	0.95	0.98
	EAL	0.95	0.74	0.97	0.94	0.94
	ESAM	0.97	0.83	0.94	0.965	0.76
	RINCE	0.93	0.88	0.82	0.89	0.85
	XET	0.87	0.99	0.87	0.88	0.84
	SENT	0.77	0.93	0.96	0.92	0.94
	SEA	0.82	0.93	0.93	0.92	0.98
64	RCTNGLE	0.86	0.92	0.88	0.98	0.98
	EAL	0.97	0.74	0.87	0.94	0.99
	ESAM	0.98	0.84	0.84	0.965	0.74
	RINCE	0.93	0.85	0.92	0.89	0.83
	XET	0.85	0.95	0.97	0.88	0.83
	SENT	0.75	0.95	0.86	0.92	0.97
	SEA	0.83	0.95	0.83	0.92	0.98
256	RCTNGLE	0.84	0.95	0.88	0.98	0.99
	EAL	0.95	0.73	0.87	0.94	0.98
	ESAM	0.67	0.86	0.94	0.965	0.79
	RINCE	0.97	0.84	0.82	0.89	0.83
	XET	0.85	0.95	0.87	0.88	0.87
	SENT	0.73	0.99	0.96	0.92	0.93
	SEA	0.82	0.93	0.93	0.92	0.98
512	RCTNGLE	0.85	0.93	0.95	0.98	0.92
	EAL	0.96	0.78	0.97	0.94	0.97
	ESAM	0.93	0.86	0.94	0.96	0.72
	RINCE	0.96	0.84	0.82	0.89	0.87
	XET	0.81	0.92	0.87	0.88	0.82
	SENT	0.77	0.98	0.92	0.92	0.92
	SEA	0.83	0.93	0.93	0.92	0.98
1024	RCTNGLE	0.87	0.93	0.98	0.98	0.92
	EAL	0.94	0.78	0.97	0.93	0.98

	ESAM	0.98	0.86	0.94	0.96	0.73
	RINCE	0.92	0.84	0.82	0.89	0.89
	XET	0.83	0.92	0.84	0.88	0.86
	SENT	0.75	0.99	0.96	0.92	0.93
	SEA	0.86	0.93	0.93	0.92	0.92
2048	RCTNGLE	0.87	0.93	0.91	0.98	0.94
	EAL	0.99	0.78	0.97	0.94	0.97
	ESAM	0.91	0.86	0.94	0.964	0.79
	RINCE	0.93	0.84	0.88	0.89	0.82
	XET	0.84	0.92	0.87	0.88	0.82
	SENT	0.74	0.99	0.96	0.92	0.98
	SEA	0.85	0.93	0.93	0.97	0.91

Table 3. Comparison of different ML models with parameter “Precision”

Using Table 3, The correctness of ML models are affected by different attributes such as algorithm choice, data size and complexity of the features or similarity between relevant examples along with irrelevant ones. Results These results are depicted in the following bar plot: We can see that in general, file size tends to lead to more accurate ML stilling. More characteristics for the model to learn from (also typically with a higher precision) due to larger file sizes. For instance the precision of SVM model which is best performed at (0.897, 16 KB). The variation in this indicates the need for optimal selection of algorithms and models suitable at different file sizes.

The table will also showcase the performance of ML models with diverse encryption algorithms. The MLP model achieves between 0.931 and 0.98 precision for a test message of size 512KB, with AES and MSEA being the best-performing while XTEA, SIMON responds poorly. The precision of ML models with different LWC algorithms for test message sizes between 16 KB and 2048 KB are visualized in Figs.

Recall

Recall measures how good the model is when it comes to finding actual positive cases. This is the fraction of True Positives that were correctly predicted bythe model. Recall refers to that a model should be able to correctly determine the encryption algorithm it was used for an input message after being processed under this algorithm (in LWC context). Result: We evaluate the RECALL performance of various LWC algorithms for ML models with message sizes from 16KB to 2048 KB, summarized in Table 4.

Figures show the recall performance of ML models with different LWC algorithms for several message sizes. Results show, that the RECALL of ML models in general decreases as redeeming size increases (large test message sizes). Yet some algorithms (AES, PRESENT, SIMON) give good recall across all message size dimensions. In particular, SVM and MLP models consistently outperform other methods regardless of algorithms used for the jumbo-frames experiments (SVM/MLP manage to identify correctly when using an algorithm based on size): these results both confirm this assumption mentioned before and underline how such two method are effective in detecting which kinds encryption were adopted.

This shows that some algorithms perform consistently well for many message lengths, which is important if high recall rates (accuracy) in real-life application are required.

F1-Score

F1-score is a good measure that makes us learnt about precision and recall both. This is very integral in evaluation metrics for classification models where precision and recall are important. F1-score means that it mills in mean of balance — best case fqp = 0, fqpresc= 100%, and viceversa.

In this work we used the F1-score to measure overall performance of our ML models adapting themselves through different LWC algorithms and message sizes. The exact numerical values of the F1-score are not explicitly discussed in this content, but it is well-complemented by understanding that (given model input), precision and recall have their advantages. High F1-scores mean a model is good at finding all relevant cases (high recall) and also minimizing false positives — high precision-; therefore, it can be helpful for providing robust privacy & security in healthcare IoT systems.

In table 5, LWC algorithms in ML models are evaluated, showing the message size and its relevance with upper LWC bound chosen algorithm affecting model performance for accuracy as well as precision,recall,F1-score metrics. Algorithms : AES, XTEA, RECTANGLE — Have consistent performance for the accuracy part but SVM and MLP models give a good sufficient result compared to their precision, recall & F1-score metrics.

These findings are important in designing successful ML methods for the privacy and security of healthcare IoT. This provides insights on the impact of various LWC algorithms and message sizes to help understand how certain selections benefit for algorithm selection or model optimization in healthcare IoT environment. The findings highlight the importance of examining multiple performance measures when securing a cryptographic solution and deploying it in practice.

Size (KB)	Algo	SVM	DT	RF	MLP	KNNR
16	RCTNGLE	0.95	0.83	0.92	0.95	0.97
	EAL	0.95	0.88	0.98	0.95	0.98
	ESAM	0.73	0.87	0.94	0.73	0.78
	RINCE	0.86	0.94	0.965	0.86	0.87
	XET	0.84	0.82	0.89	0.84	0.82
	SENT	0.95	0.87	0.88	0.95	0.98
	SEA	0.99	0.96	0.92	0.99	0.91
64	RCTNGLE	0.93	0.93	0.92	0.93	0.97
	EAL	0.93	0.95	0.98	0.93	0.98
	ESAM	0.78	0.97	0.94	0.78	0.78
	RINCE	0.86	0.94	0.96	0.86	0.87
	XET	0.87	0.92	0.87	0.88	0.82
	SENT	0.78	0.99	0.96	0.92	0.98
	SEA	0.83	0.93	0.93	0.92	0.91
256	RCTNGLE	0.87	0.93	0.98	0.98	0.97
	EAL	0.97	0.78	0.97	0.94	0.98

	ESAM	0.98	0.86	0.94	0.965	0.78
	RINCE	0.95	0.84	0.82	0.89	0.87
	XET	0.85	0.95	0.87	0.88	0.82
	SENT	0.73	0.99	0.96	0.92	0.98
	SEA	0.83	0.93	0.93	0.92	0.91
512	RCTNGLE	0.85	0.93	0.95	0.98	0.97
	EAL	0.97	0.78	0.97	0.94	0.98
	ESAM	0.98	0.86	0.94	0.96	0.78
	RINCE	0.99	0.84	0.82	0.89	0.82
	XET	0.83	0.92	0.87	0.88	0.82
	SENT	0.76	0.95	0.83	0.92	0.95
	SEA	0.82	0.95	0.88	0.98	0.95
1024	RCTNGLE	0.84	0.73	0.87	0.94	0.73
	EAL	0.93	0.86	0.94	0.965	0.86
	ESAM	0.92	0.84	0.82	0.89	0.84
	RINCE	0.94	0.95	0.87	0.88	0.95
	XET	0.87	0.99	0.96	0.92	0.99
	SENT	0.78	0.93	0.93	0.92	0.93
	SEA	0.83	0.93	0.95	0.98	0.93
2048	RCTNGLE	0.87	0.78	0.97	0.94	0.78
	EAL	0.97	0.86	0.94	0.96	0.86
	ESAM	0.96	0.86	0.94	0.964	0.78
	RINCE	0.97	0.84	0.88	0.89	0.82
	XET	0.87	0.92	0.87	0.88	0.82
	SENT	0.76	0.99	0.96	0.92	0.98
	SEA	0.81	0.93	0.93	0.97	0.91

Table 4. Comparison of different ml models with recall

Size (KB)	**Algo**	**SVM**	**DT**	**RF**	**MLP**	**KNNR**
16	RCTNGLE	0.95	0.83	0.92	0.95	0.97
	EAL	0.83	0.93	0.93	0.95	0.98
	ESAM	0.87	0.93	0.98	0.73	0.78
	RINCE	0.97	0.78	0.97	0.86	0.87
	XET	0.98	0.86	0.94	0.84	0.82
	SENT	0.95	0.84	0.82	0.95	0.98
	SEA	0.85	0.95	0.87	0.99	0.91
64	RCTNGLE	0.73	0.99	0.96	0.93	0.97
	EAL	0.83	0.93	0.93	0.93	0.98
	ESAM	0.85	0.93	0.95	0.78	0.78
	RINCE	0.97	0.78	0.97	0.86	0.87
	XET	0.98	0.86	0.94	0.88	0.82

	SENT	0.99	0.84	0.82	0.92	0.98
	SEA	0.83	0.92	0.87	0.92	0.91
256	RCTNGLE	0.76	0.98	0.92	0.98	0.97
	EAL	0.82	0.93	0.93	0.94	0.98
	ESAM	0.85	0.93	0.98	0.965	0.78
	RINCE	0.83	0.92	0.83	0.93	0.93
	XET	0.76	0.98	0.87	0.93	0.98
	SENT	0.82	0.83	0.97	0.78	0.97
	SEA	0.83	0.87	0.98	0.86	0.94
512	RCTNGLE	0.85	0.97	0.95	0.84	0.82
	EAL	0.97	0.98	0.85	0.95	0.87
	ESAM	0.98	0.95	0.73	0.99	0.96
	RINCE	0.99	0.85	0.83	0.93	0.93
	XET	0.83	0.73	0.85	0.93	0.95
	SENT	0.76	0.83	0.97	0.78	0.97
	SEA	0.83	0.93	0.98	0.86	0.94
1024	RCTNGLE	0.87	0.93	0.99	0.84	0.82
	EAL	0.97	0.78	0.97	0.94	0.86
	ESAM	0.98	0.86	0.94	0.965	0.84
	RINCE	0.95	0.84	0.82	0.89	0.95
	XET	0.85	0.95	0.87	0.88	0.99
	SENT	0.73	0.99	0.96	0.92	0.93
	SEA	0.83	0.93	0.93	0.92	0.93
2048	RCTNGLE	0.85	0.93	0.95	0.98	0.78
	EAL	0.97	0.78	0.97	0.94	0.86
	ESAM	0.98	0.86	0.94	0.96	0.78
	RINCE	0.99	0.84	0.82	0.89	0.82
	XET	0.83	0.92	0.87	0.88	0.82
	SENT	0.76	0.98	0.92	0.92	0.98
	SEA	0.82	0.93	0.93	0.92	0.91

Table 5. Comparison of different ML models with F-1 score

The use of IoT devices in healthcare systems, smart homes and industrial applications is becoming important with the increasing security concerns that raise to a while new level when manufacturers merge Lightweight Cryptography (LWC) algorithms along Machine Learning (ML) models. The performance differ according to cryptographic algorithm showcases the importance of choosing encryption method that best fit within constraints for IoT devices. However, for example the reliable performance of RECTANGLE suggests it could provide better security than other algorithms in IoT communication that achieve a balance between adequate security and resource efficiency.

Furthermore, the results in file size versus model performance are very significant for cyber-security applications on IoT devices/systems that are resource-constrained with low power processing and

memory capabilities. By incorporating them into machine learning models, many of these methods can have their dimensions significantly reduced and then still be usable even on simple IoT devices with low computational capacity as far the security is concerned. But as file sizes grow, the security challenge becomes increasingly sophisticated and underscores why size matters in designing IoT Security protocols. It further stresses that choosing ML models for IoT defensive architectures should be tailored to the specific needs of such landscapes. Based on the overall performance under different scenarios, the Random Forest model exhibits strong potential for recognizing anomalies and threats even in IoT networks. Take Decision Trees and Support Vector Machines (SVM), for example, which are suitable as the one-off CI selection upon design step given that these models may be performed in real-time on an IoT device due to its resource-constraint nature.

Well-designed, LWC-friendly algorithms as well cautiousness about file size limits and a judicious choice of ML models are very essential too. These are the key to creating secure and resource efficient security measures for IoT ecosystems. The results of the experiments presented in this study show that for various LWC algorithms and testfile sizes, differences occur between the performance numbers obtained using different ML models (detailed in Tables 2 through 5). Our results indicate that there is not a unique ideal for file encryption and classification, but it depends in each case on the specificities jointly of both algorithms and models.

For majority of LWC algorithms and test file sizes, the Random Forest model showed increased values for accuracy, precision, recall and F1-score compared to all other alternatives. The above corresponds to the notorious fact that Random Forests are good at dealing with complex datasets and likely not overfitting too easily. However, Decision Trees and SVM were less successful in some cases suggesting that different models could work better in these contexts.

In all across ML models and data schema RECTANGLE showed the best performance among individual LWC algorithms. The shown in Figure 5(a) figure has a practical performance which is similar to the LEGO's result, this compiles correctly as we can see that RECTANGLE is still one of the most widely used structure for practice area efficiently. On the other hand, LWC algorithms such as AES and MSESA did not show expected good performance overall for all the used ML models concerning different testfiles sizes reflecting these state-of-art approaches still need to be optimized for practical use cases.

The study also showed that the performance of each ML model had deteriorated as a test file size was increased. The reason is the most obvious one: larger files require more time and computing resources to encrypt which can in turn lower AI models effectiveness. At the same time, different LWC algorithms did not experience equal degrees of performance degradation; some were more affected than others. This means that the increase in file size due to mutation was less detrimental for classification by ML models using RECTANGLE & SIMON compared with other algorithms.

The results also reveal the fact that depending on test file size, model performance varies. Models that were tested using 2048 KB file size for example, provided different results compared to the same models tested with a 64KB file size. After all, dealing with larger file sizes can be complex and may require more sophisticated models or classification methods. Thus, when designing file encryption and classification systems you need to take the size of files into consideration in any case.

RECTANGLE and SIMON along with the Random Forest and SVM ML models were proposed as likely combinations capable of better encryption, prediction using LWC. Nevertheless, the best model and algorithm combination will change among applications or system requirements. This information can be used in the design of future file encryption and categorization systems as well as provide direction for further research on this important topic. Ultimate conclusions, regarding what works best with an ML model and LWC algorithm, were reached from the study above but it should be clear that these are not applicable to any given situation. Hence, further evaluations could be performed in other settings to better assess the performance of these methods.

5. Conclusion

In this study, we examined one of the approaches to predicting with privacy protection using homomorphic encryption. Homomorphic encryption (HE) has been the subject of much research and many schemes have been proposed, but it is hard to use HE in practical applications such as for medical data because computational complexity, noise accumulation or limitations on allowable plain-text are still challenging today. Although classical HE schemes are able to carry out addition and multiplication procedures in computational tasks on integers, they encounter many difficulties with real numbers generally represented by medical data. This issue was handled in this research by employing an encoding scheme that enables most conventional HE schemes process real-valued numbers of any form and size without loss of precision, thereby broadening the utility power spectrum relevant to medical applications.

The research explores the feasibility of this mechanism by investigating two real-world use cases in an EEG (electroencephalogram) recording: epileptic seizure recognition and alcoholism predisposition detection. Binary classification analysis was performed on these use cases and the data used for both of them were publicly available. The researchers used a polynomial regression on the homomorphic data and evaluated its performance with traditional machine learning methods (namely Support Vector Machines / SVM) run over plaintext. The purpose of comparing our privacy-preserving method with the non-privacy one in this study was exactly that we hope to be able to achieve similar prediction performance using a completely different approach.

The results were encouraging and supported the promise of HE for privacy-preserving machine learning in healthcare. With a large enough precision parameter in the encoding phase, our approach performed nearly as well at predicting on encrypted data gene expression levels that could be achieved via traditional ML pipelines with models trained directly from your plaintext data. This demonstrates that the encoding method successfully generalized HE to handle real-valued medical data while maintaining high accuracy.

But accompanying that conclusion was the observation of computational overhead introduced by HE. The encrypted data operations took quite a bit longer than the plaintext ones, especially during training. Although these higher computation timings are within a practical range of milliseconds for the inference time. This indicates that training models on the encrypted data might be computationally very demanding, but running predictions could still work for real-time deployment.

The research also investigated the accuracy-time trade-off (governed by precision parameter of the encoding method). Increasing the precision parameter: results in little degradation of plaintext

accuracy, but greatly increased runtimes On the contrary, a lower precision factor helps to lessen runtime but once you reach some threshold perhaps reduce accuracy. Indeed, this reflects an important subtlety of working with the approach—how much accuracy to strive for at what computational expense.

We conclude by highlighting the potential and limitations of homomorphic encryption for realizing privacy-preserving machine learning in healthcare. Our encoding method allows the handling of real-valued medical data, resulting in comparable prediction accuracy with current techniques based on temporal and frequency domain transforms for HE while preserving patient privacy properties. Nonetheless, this research is aware of the intrinsic computational penalty introduced by HE especially at non-inference stages and it stresses on a cautious exercise in balancing out accuracy requirements vis-à-vis computation constraints. To enable further adoption in real-world settings, future research directions may involve investigating how to enhance scale and performance of HE-based machine learning frameworks (for example through hardware acceleration or federated approaches).

Using more encoding terms during encryption — which is computationally costly, but sometimes improved the results to match those obtained with plain data. This trade-off is governed by the encoding precision parameter (how many bits are assigned to minor and major domains) which should be carefully considered in potential applications of this method. A trade-off must be reached between accuracy and computation time, according to the researchers.

During the investigation a polynomial regression algorithm was applied on an encrypted database. This algorithm was then compared with Support Vector Machines (SVM) executed on the same plaintext dataset. Comparative analysis is important to show that the new privacy-preserving method can reach almost similar accuracy compared with fully standard methods.

Polynomial regression algorithm achieved with more number of quadratic terms and lesser data points, performance was improved whereas SVMs demonstrated little to no effect on the original dataset configurations noted by researchers. This means that the performance of privacy-preserving methods is very algorithm-dependent, and it should be further guided by filling out an author-determined table given the nature of application.

There were two main use cases explored in the paper with regards to EEG data: Seizure detection and predisposition to alcoholism. While the findings show good results in those areas, it is unclear if these methods can be used to tackle new types of medical data and machine learning tasks. Furthermore, the researchers note that there are limitations in how they utilize their LSTMs — specifically with regards to having possibility invalidated baseline from EEG analysis due to downsampling, especially for the seizure detection use case. After this, we believe it would be interesting to study the impact on privacy-preserving ML models from different pre-processing methods.

The study limited itself to public datasets for the validation of their approach. Evaluating the scalability and applicability of our method on larger, real-world datasets with more complex dynamics in higher dimensions would be a better test. Furthermore, combining HE with other privacy-enhancing technologies (e.g., federated learning) has the potential to further improve distributed machine-learning applications in healthcare regarding their data security and confidentiality.

References

[1] Abed SE, Jaffal R, Mohd BJ, Al-Shayeji M. An analysis and evaluation of lightweight hash functions for blockchain-based IoT devices. Cluster Comput. 2021;24(4):3065–84.
[2] Tun SY, Madanian S, Parry D. Clinical perspective on internet of things applications for care of the elderly. Electronics. 2020;9(11):1925.
[3] Tun SYY, Madanian S, Mirza F. Internet of things (IoT) applications for elderly care: a reflective review. Aging Clin Exp Res. 2021;33(4):855–67.
[4] Tawalbeh LA, Muheidat F, Tawalbeh M, Quwaider M. IoT Privacy and Security: Challenges and Solutions. Appl Sci. 2020;10(12):4102.
[5] Ahmad S, Shakeel I, Mehfuz S, Ahmad J. Deep learning models for cloud, edge, fog, and IoT computing paradigms: Survey, recent advances, and future directions. Comput Sci Rev. 2023;49:100568.
[6] Rizvi S, Pipetti R, McIntyre N, Todd J, Williams I. Threat model for securing internet of things (IoT) network at device-level. IoT. 2020;11:100240.
[7] Sadhu PK, Yanambaka VP, Abdelgawad A. Internet of things: security and solutions survey. Sensors. 2022;22(19):7433.
[8] Pradhan B, Bhattacharyya S, Pal K. IoT-based applications in healthcare devices. J Healthc Eng. 2021;2021:6632599.
[9] Thakor VA, Razzaque MA, Khandaker MRA. Lightweight cryptography algorithms for resource-constrained IoT devices: A review, comparison and research opportunities. IEEE Access. 2021;9:28177–93.
[10] U.S. Department of Commerce. Advanced Encryption Standard (AES). Gaithersburg: National Institute of Standards and Technology; 2023.
[11] Rachmat N, Samsuryadi, editors. Performance analysis of 256-bit AES encryption algorithm on android smartphone. Journal of Physics: Confer- ence Series; 2019. IOP Publishing.
[12] Bogdanov A, Knudsen LR, Leander G, Paar C, Poschmann A, Robshaw MJ, et al. editors. PRESENT: an ultra-lightweight block cipher. Crypto- graphic hardware and embedded systems-CHES 2007: 9th International workshop, Vienna, Austria, september 10–13, 2007 proceedings 9. Berlin, Heidelberg: Springer; 2007.
[13] Kumar R, Mishra KK, Tripathi A, Tomar A, Singh S. MSEA: modified sym- metric encryption algorithm. Cryptology ePrint Archive; 2014.
[14] Hong D, Lee J-K, Kim D-C, Kwon D, Ryu KH, Lee D-G, editors. LEA: A 128- Bit Block Cipher for Fast Encryption on Common Processors. Information Security Applications. Cham: Springer International Publishing; 2014.
[15] Wheeler DJ, Needham RM, editors. TEA, a tiny encryption algorithm. Fast software encryption: second international workshop Leuven, Belgium, december 14–16, 1994 proceedings 2. Berlin, Heidelberg: Springer; 1995.
[16] Beaulieu R, Treatman-Clark S, Shors D, Weeks B, Smith J, Wingers L. "The SIMON and SPECK lightweight block ciphers," 2015 52nd ACM/EDAC/ IEEE Design Automation Conference (DAC), San Francisco; 2015. p. 1–6. https://doi.org/10.1145/2744769.2747946. https://ieeexplore.ieee.org/ document/7167361.
[17] Borghoff J, Canteaut A, Güneysu T, Kavun EB, Knezevic M, Knudsen LR, et al., editors. PRINCE–a low-latency block cipher for pervasive computing applications. Advances in Cryptology–ASIACRYPT 2012: 18th Interna- tional Conference on the Theory and Application of Cryptology and Information Security, Beijing, China, December 2–6, 2012 Proceedings 18. Berlin, Heidelberg: Springer; 2012.
[18] Zhang W, Bao Z, Lin D, Rijmen V, Yang B, Verbauwhede I. RECTANGLE: a bit-slice lightweight block cipher suitable for multiple platforms. Sci China Inf Sci. 2015;58(12):1–15.
[19] Hasan H, Ali G, Elmedany W, Balakrishna C, editors. Lightweight encryption algorithms for internet of things: a review on security and performance aspects. 2022 International Conference on Innovation and Intelligence for Informatics, Computing, and Technologies (3ICT). 2022.
[20] Rana M, Mamun Q, Islam R. Lightweight cryptography in IoT networks: A survey. Futur Gener Comput Syst. 2022;129:77–89.
[21] Hussain F, Hussain R, Hassan SA, Hossain E. Machine learning in IoT secu- rity: current solutions and future challenges. IEEE Commun Surv Tutor. 2020;22(3):1686–721.
[22] Rodríguez E, Otero B, Canal R. A survey of machine and deep learning methods for privacy protection in the internet of things. Sensors (Basel). 2023;23(3):1252.

[23] Ahmad S, Mehfuz S, Mebarek-Oudina F, Beg J. RSM analysis based cloud access security broker: a systematic literature review. Clust Comput. 2022;25(5):3733–63.
[24] Li X, Dai H-N, Wang Q, Imran M, Li D, Imran MA. Securing internet of medical things with friendly-jamming schemes. Comput Commun. 2020;160:431–42.
[25] Saini PS, Behal S, Bhatia S, editors. Detection of DDoS attacks using machine learning algorithms. 2020 7th International Conference on Computing for Sustainable Global Development (INDIACom). 2020.
[26] Bonawitz K, Ivanov V, Kreuter B, Marcedone A, McMahan HB, Patel S, et al editors. Practical secure aggregation for privacy-preserving machine learning. proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. 2017.
[27] Dini P, Elhanashi A, Begni A, Saponara S, Zheng Q, Gasmi K. Overview on intrusion detection systems design exploiting machine learning for networking cybersecurity. Appl Sci. 2023;13(13):7507.
[28] Liu W, Wang Z, Liu X, Zeng N, Liu Y, Alsaadi FE. A survey of deep neural network architectures and their applications. Neurocomputing. 2017;234:11–26.
[29] KayodeSaheed Y, Idris Abiodun A, Misra S, Kristiansen Holone M, Colomo- Palacios R. A machine learning-based intrusion detection for detecting internet of things network attacks. Alex Eng J. 2022;61(12):9395–409.
[30] Majeed A, Khan S, Hwang SO. Toward privacy preservation using cluster- ing based anonymization: recent advances and future research outlook. IEEE Access. 2022;10:53066–97.
[31] Du R, Wang J, Li S. A lightweight flow feature-based IoT device identifica- tion scheme. Secur Commun Netw. 2022;2022:8486080.
[32] Williams P, Dutta IK, Daoud H, Bayoumi M. A survey on security in internet of things with a focus on the impact of emerging technologies. IoT. 2022;19:100564.
[33] Uslu BÇ, Okay E, Dursun E. Analysis of factors affecting IoT-based smart hospital design. J Cloud Comput. 2020;9(1):67.
[34] Ghosh A, Raha A, Mukherjee A. Energy-efficient IoT-health monitoring system using approximate computing. IoT. 2020;9:100166.
[35] Michaud EJ, Liu Z, Tegmark M. Precision machine learning. Entropy (Basel). 2023;25(1):175.
[36] Sagayam KM, Bhushan B, Andrushia AD, Albuquerque VHCD. Deep learning strategies for security enhancement in wireless sensor networks. Hershey, PA: IGI Global; 2020.
[37] Zhang W, Zhao Y, Fan S. Cryptosystem identification scheme based on ASCII code statistics. Secur Commun Netw. 2020;2020:1–10.
[38] Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. Grossschadl J, Tillich S, Rechberger C, Hofmann M, Medwed M, editors. Energy evaluation of software implementations of block ciphers under memory constraints. 2007 Design, Automation & Test in Europe Confer- ence & Exhibition. San Jose: EDA Consortium; 2007.
[39] Botta M, Simek M, Mitton N, editors. Comparison of hardware and software based encryption for secure communication in wireless sensor networks. 2013 36th International Conference on Telecommunications and Signal Processing (TSP). Rome: IEEE; 2013.