

# Machine Learning in Precision Medicine to Preserve Privacy via Encryption

William Briguglio<sup>a</sup>, Parisa Moghaddam<sup>a</sup>, Waleed A. Yousef<sup>a,b,\*</sup>, Issa Traoré<sup>a</sup>, Mohammad Mamun<sup>c</sup>

<sup>a</sup>ECE dep., University of Victoria, Victoria, BC, Canada.

<sup>b</sup>CS dep., Helwan University, Cairo, Egypt.

<sup>c</sup>National Research Council of Canada, NB, Fredericton, Canada.

## Abstract

Precision medicine is an emerging approach for disease treatment and prevention that delivers personalized care to individual patients by considering their genetic makeups, medical histories, environments, and lifestyles. Despite the rapid advancement of precision medicine and its considerable promise, several underlying technological challenges remain unsolved. One such challenge of great importance is the security and privacy of precision health-related data, such as genomic data and electronic health records, which stifle collaboration and hamper the full potential of machine-learning (ML) algorithms. To preserve data privacy while providing ML solutions, this article makes three contributions. First, we propose a generic machine learning with encryption (MLE) framework, which we used to build an ML model that predicts cancer from one of the most recent comprehensive genomics datasets in the field. Second, our framework's prediction accuracy is slightly higher than that of the most recent studies conducted on the same dataset, yet it maintains the privacy of the patients' genomic data. Third, to facilitate the validation, reproduction, and extension of this work, we provide an open-source repository that contains the design and implementation of the framework, all the ML experiments and code, and the final predictive model deployed to a free cloud service.

**Keywords:** Machine Learning, Encryption, Homomorphic Encryption, Precision Medicine, Privacy

## 1. Introduction

Precision medicine is a departure from one-size-fits-all medicine toward the customization of disease treatment and prevention for individuals by leveraging their individual variability in genes, environment, and lifestyle. Two key aspects of the digital revolution of the last decade have been an exponential increase in the amount of data being generated and a parallel increase in hardware and GPU performance. While significant progress has been made in storing huge amounts of data via big-data technologies, the ability to develop actionable knowledge that facilitates individualized care through precision medicine has lagged behind.

Despite the exciting prospects of precision healthcare, it faces several technical and societal hurdles related to the identification of health risks, diagnoses, and outcomes by analyzing data extracted from integrated biomedical databases. Security and privacy concerns are such hurdles. While precision health provides tremendous benefits by enabling better care, it can lead to personal privacy breaches through genetic disclosure or genetic discrimination. To deliver targeted, personalized care, personal data (e.g., specific human genome

sequencing) must be shared with many professionals in possibly diverse geographic locations or jurisdictions and sometimes over unreliable channels, such as the internet. This poses several risks, such as insider threats, social engineering, distributed denial of service (DDoS), illicit data inferences, cyber bullying/blackmailing, etc. (Rahman et al., 2019). Unlike protected health information (PHI), precision health data, such as genomic data, not only identifies patients but also multiple generations of their families. Hence, such data can be leveraged to conduct targeted security and privacy attacks against vulnerable individuals or groups of related individuals if fallen in the hands of malicious actors.

In principle, a machine-learning (ML) model can be trained on either confidential or public data, allowing more training samples, data distributions, and therefore more complex, predictive, and generalizing models. These complex models can theoretically achieve higher predictive performance and find novel associations within precision healthcare. However, performing analytics on new cases provided by hospitals or medical centers should be treated with the utmost privacy preservation level for the reasons introduced above.

From this perspective, the purpose of the present article is threefold. First, we propose a machine learning with encryption (MLE) framework that considers these requirements and constraints and facilitates performing analysis on a real precision healthcare dataset while preserving its privacy. Second, we illustrate the model's success by considering a case study using the MSK-IMPACT dataset, one of the most recent com-

\* Corresponding Author

Email addresses: wbriguglio@uvic.ca (William Briguglio), ParisaMoghaddam@uvic.ca (Parisa Moghaddam), wyousef@UVIC.ca, wyousef@fci.Helwan.edu.eg (Waleed A. Yousef), itraore@ece.uvic.ca (Issa Traoré), Mohammad.Mamun@nrc-cnrc.gc.ca (Mohammad Mamun)

prehensive genomic datasets in the field (Zehir et al., 2017), and we produce a predictive model for cancer with higher accuracy than the most recent publications on the same dataset (Penson et al., 2020); while preserving the privacy of the cases as required. Third, to facilitate the capabilities of different communities, such as software engineering, ML, and precision medicine, and for the extension and validation of this work, we provide the following as an open-source repository: (1) the system design and implementation of the MLE framework, (2) all the ML code and experiments on the MSK-IMPACT dataset, and (3) a free cloud service for medical practitioners to predict their own cases using the MLE framework.

The remainder of the present article is as follows. Section 2 introduces homomorphic encryption (HE) and related work on the joint field of ML and HE. Section 3 presents our proposed MLE framework and its main modules. Section 4 introduces the MSK-IMPACT dataset as a case study, describes the ML experiments that we conducted on this dataset to provide a predictive model with privacy preservation via HE, and introduces the open-source repository for this work. Section 5 concludes the article and suggests future work that would complement the current work. For the present article to be self-contained, the online supplementary material provides more explanation of the basics of genomics necessary for our article as well as an example of HE.

## 2. Related Work and Literature Review

Any privacy-preserving precision health analytics builds on two main components: (1) the security and privacy features required to protect interactions with the data by stakeholders and (2) ML predictive models for this data. Each is reviewed briefly below.

### 2.1. Homomorphic Encryption

Encryption is the process of converting data from something intelligible into something unintelligible by sealing data in a metaphorical vault that can only be opened by somebody holding the secret decryption key to prevent unauthorized personnel from viewing them. A special scheme of encryption is HE, which was originally proposed by Rivest et al. (1978) as a way to encrypt data such that certain operations could be performed on them without possessing that secret key (i.e., without decryption). The term “homomorphic encryption” describes a class of encryption algorithms that satisfies the homomorphic property; that is, certain operations, such as addition, can be carried out on ciphertext directly so that upon decryption the same answer is obtained as operating on the original messages. Therefore, HE allows other parties (e.g., the cloud and service providers) to calculate certain mathematical functions expressed only in terms of these operations on the encrypted data while preserving the function and format of the encrypted data. For brevity, these types of functions are referred to as “HE-friendly”. Formally, this can

be expressed as

$$Dec[k_s, Enc(k_p, m_1) \diamond Enc(k_p, m_2)] = m_1 \circ m_2, \quad (1)$$

where  $k_s, k_p$  are the secret and public keys, respectively (since they are not equal, this is called “asymmetric encryption”),  $m_1, m_2 \in M$  are two values on which we wish to perform encrypted operations on,  $M$  is the message space of the HE scheme (i.e., the set of all possible values acceptable by the scheme), and  $\diamond, \circ$  are operations in encrypted and plain-text space, respectively. Like other types of encryption schemes, HE has three main functions: key generation, message encoding, and decoding (the supplementary material provides some details on these operations.)

The remarkable properties of HE schemes are not without limitations. First, the set of functions that can be computed in ciphertext space is very restricted. Second, the computational complexity of the HE scheme depends primarily on the level of multiplication (i.e., the degree of the polynomial being evaluated) carried out on the encrypted data. Third, the ciphertext size increases considerably after encryption. Fourth, random noise is added to encrypted values for security reasons that varies with the type of operation (e.g., multiplication increases noise much more than addition). If this noise grows too large, then decryption yields incorrect results. There are three basic approaches to implement HE (see, e.g., Acar et al., 2018):

**partially homomorphic encryption (PHE)**, which allows only one type of operation to be executed on encrypted values an unlimited number of times.

**somewhat homomorphic encryption (SWHE)**, where the size of the ciphertext grows with each homomorphic operation and hence the maximum number of allowed homomorphic operations is limited.

**fully homomorphic encryption (FHE)**, which supports an unlimited number of additions and multiplications (Gentry, 2009). This property makes FHE the most sophisticated HE scheme and the “holy grail” of modern cryptography. The FHE scheme supports basic arithmetic computations on encrypted data. Despite being a potential cryptographic technique, however, some FHE schemes remain impractical for real-world applications due to their computational overhead.

### 2.2. Machine Learning with Encryption

Any ML algorithm trains on some training dataset  $\mathbf{tr}$ , fits a model’s parameters  $\mathcal{M}$ , and finally tests on a testing dataset  $\mathbf{ts}$ . Therefore, in principle, there are eight possible combinations or scenarios to introduce privacy via encryption to the learning process by encrypting (or leaving unencrypted) each of these three components. Table 1 summarizes those eight scenarios; below, we provide more details on them and their connections to the present article. We use 0 to denote an unencrypted component, where it still can only be encrypted using the public key  $k_p$  of another component without having access to its private key  $k_s$ , and we use 1 to denote an en-

Table 1: The eight possible scenarios of encrypting the three components: the training dataset  $\mathbf{tr}$ , the ML model parameters  $\mathcal{M}$ , and the testing dataset  $\mathbf{ts}$ .

#	$\mathbf{tr}$	$\mathcal{M}$	$\mathbf{ts}$	Literature	Dataset	ML	Enc. Library
0	0	0	0	Ordinary ML			
1	0	0	1	Our present article Dowlin et al. (2016) Hesamifard et al. (2017)	MSK MINIST MINIST, CIFAR-10	many NN DNN	SEAL SEAL Helib
2	0	1	0	Bost et al. (2015)	Multiple	NB, HP, DT	self-implementation
3	0	1	1	—			
4	1	0	0	Graepel et al. (2013b) Aslett et al. (2015a) Nandakumar (2019)	Wisconsin Multiple MNIST6	FDA NB, RF DNN7	Magma EncryptedStats HElib
5	1	0	1	—	Not possible under the current theory		
6	1	1	0	—	Not practical: training on encrypted data already produces encrypted model		
7	1	1	1	—	Not practical: training on encrypted data already produces encrypted model		

encrypted component, where its private key  $k_s$  is not available for the other two components.

Scenario 0 is denoted by the binary combination 000, when  $\mathbf{tr}$ ,  $\mathcal{M}$ , and  $\mathbf{ts}$  are all not encrypted; this is the typical ML paradigm, where no privacy is of concern.

Scenario 1 is denoted by 001, where only  $\mathbf{ts}$  needs to be encrypted; this is the scenario of the present article, as Sections 3 and 4 describe. In such a scenario, since the model has access to an unencrypted training set, such as a public dataset, only the client data  $\mathbf{ts}$ , which could be patients’ genome records, are sensitive. Although the standard homomorphic property as defined in (1) would imply that  $\mathcal{M}$  must be encrypted with  $k_p$  to predict on an encrypted  $\mathbf{ts}$ , this is not the case for our work, which leverages special techniques implemented in SEAL Microsoft (2020), that allow encryption with plain-text multiplication with the caveat that the results themselves are encrypted and can only be decrypted with  $k_s$ . Research exists in this category but in areas of application other than precision medicine. Dowlin et al. (2016) showed that a cloud service is capable of applying a neural network (NN) to encrypted  $\mathbf{ts}$  to make encrypted predictions and return them in encrypted forms. They constructed a convolution NN (CNN) model from the unencrypted MINIST dataset and then produced a simpler FHE-friendly version of the CNN constructed only from addition and multiplication operations so that the parameters could be encrypted using the public key of the private testing dataset  $\mathbf{ts}$ . Hesamifard et al. (2017) developed new techniques to allow testing CNN on encrypted  $\mathbf{ts}$ . First, they designed methods to approximate the activation functions commonly used in CNNs with low-degree, FHE-friendly polynomials. Then, they trained a CNN on unencrypted  $\mathbf{tr}$  with the approximation polynomials instead of the original activation functions. Finally, they converted the trained CNN to make predictions on encrypted  $\mathbf{ts}$ .

Scenario 2 is denoted by 010, where the model is trained on an unencrypted training dataset  $\mathbf{tr}$ . However, the model parameters themselves are then encrypted, which may imply privacy in  $\mathbf{tr}$ , as well if the training is pursued locally where  $\mathbf{tr}$  resides. Although the testing data  $\mathbf{ts}$  is denoted by 0, it must be sent to  $\mathcal{M}$  encrypted with its public key, as it is not possible, according to the theory of FHE, to pursue binary op-

erations on encrypted numbers (parameters of  $\mathcal{M}$ ) and unencrypted numbers ( $\mathbf{ts}$ ), without the results being encrypted and only decryptable with the same  $k_s$  that can decrypt  $\mathcal{M}$ . The virtue of scenario 2 is that it entails more freedom in choosing the model  $\mathcal{M}$  as opposed to scenarios 4–7, where  $\mathbf{tr}$  is encrypted and a stringent limitation is incurred for choosing the model  $\mathcal{M}$  that can train on encrypted data. We are not aware of any literature that applies scenario 2 explicitly; however, Bost et al. (2015) provided a very nested, layered model that could be classified as 010 scenario, but without relying solely on HE. They implemented a decision tree, naive Bayes and hyperplane decision that could test (not train) on encrypted data and built their models using cryptographic “building blocks” that emphasized protecting the model parameters and test data. They also used garbled circuits to compare encrypted data, which allowed a construction of argmax with alterations to ensure the ordering was not leaked. However, this introduced  $k-1$  round trips between the party performing argmax and the party that can perform decryption. These building blocks allowed the implementation of decision tree, naive Bayes, and hyperplane decision with some minor changes. The building blocks also allowed the construction of other ML methods and a combination of methods using AdaBoost, which the authors demonstrated.

Scenario 3 is like 2 in that the model is trained on an unencrypted  $\mathbf{tr}$ , and  $\mathcal{M}$ ’s parameters are then encrypted; however, the test dataset  $\mathbf{ts}$  is also encrypted with a different  $k_s$  than  $\mathcal{M}$ . Since there is no known method in the literature that allows the use of binary operations on two numbers encrypted with different  $k_s$ , scenario 3 (011) is not theoretically feasible under the current theory of cryptography.

Scenario 4 is denoted by 100, where an ML model is trained on encrypted  $\mathbf{tr}$  (as in scenarios 5–7, as well). Hence, the model  $\mathcal{M}$  will have encrypted weights by product, and the testing data must be sent to  $\mathcal{M}$  encrypted with the same public key of  $\mathbf{tr}$ , as explained above. Therefore, the reason scenario 5 (101) is not theoretically possible is the same as scenario 3. Furthermore, scenarios 6 and 7 (11x) are not of any practical interest, since the produced encrypted  $\mathcal{M}$  does not need further encryption; this is possibly the reason for the absence of literature on these two scenarios. Under scenario 4, Graepel et al. (2013a) defined a fully confidential version of linear means and Fisher’s discriminant analysis (FDA), which can train and test on encrypted data. Linear means are rewritten to avoid division when learning the weights. The resulting decision function returns a multiple of the original decision function with the same sign. However, FDA requires the inverse of the covariance matrix to obtain the feature weights. This is found using gradient descent, the  $r^{th}$  iteration of which is shown to be a  $d$ -degree polynomial, where  $d = 2(r-1) + 1$ . Aslett et al. (2015a) provided a completely random forest (CRF) implementation that could train and test on encrypted data. Among other alterations to the algorithm, the key difference was encoding feature values using one hot encoding after quantile partitioning. CRFs have important benefits, especially learning incrementally. The authors also provided a naive Bayes classifier that could train and test on encrypted

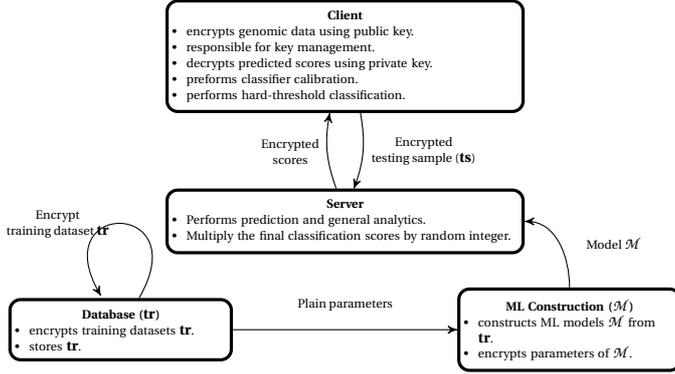


Figure 1: A block diagram of a privacy-preserving MLE framework for precision medicine that can accommodate any of the eight scenarios of Table 1. Each block contains the functionalities that can be performed in that block depending on the adopted scenario.

data. It avoided parametric Gaussian modeling of predictors by directly modeling the decision boundary  $x_j m_j + b_j$  for each predictor  $X_j$ . This required a homomorphic implementation of regression, which they also provided. Hesamifard et al. (2017) discussed the computational complexity that makes NN training on encrypted data impractical despite being theoretically possible. Other authors have targeted simpler ML algorithms to avoid heavy computations of encrypted NNs. However, Dowlin et al. (2016) show that training CNNs on encrypted data is possible. If all the activation functions and the loss function are polynomials, back-propagation can be computed using only addition and multiplication. However, high-degree polynomials used during back-propagation make it computationally prohibitive. Nandakumar (2019) evaluated the feasibility of training NNs on encrypted data completely non-interactively. His proposed system used the FHE toolkit HElib to implement stochastic gradient descent (SGD) for training. He used “ciphertext packing” to minimize the number of required bootstrapping operations and to enable the parallelization of computations at each neuron, thereby significantly reducing the computational complexity. This, in combination with simplifying the network architecture, allowed him to practically train neural networks over encrypted data despite the computational hurdles.

### 3. A Machine Learning with Encryption (MLE) Framework

In this section we propose a simple four-component system architecture to accommodate any of the eight scenarios of encryption in Table 1. This simple architecture, illustrated in Figure 1, fulfills the privacy-preserving requirements that are mandatory for future ML-based precision medicine. The components of this architecture are explained below.

**Database (tr)** is a reservoir for both publicly available genetic datasets, which do not require preserving privacy, and private datasets, which need encryption prior to public sharing. Whenever a new dataset is revealed, it can be added to this reservoir for more accurate future analytics.

**ML Construction ( $\mathcal{M}$ )** is the engine that constructs models—including transformation, feature selection, resampling, etc.—from the datasets in the *database* module. This module can be open-sourced for the entire community and can always be updated as new ML methods merge or more accurate models are constructed. In addition, the module can train on its own private dataset, which is not part of the *database* module, and then encrypt its model parameters  $\mathcal{M}$ . Alternatively, it can establish a protocol with the *database* module to train on the private dataset without encryption for a wider range of algorithms and then encrypt the model parameters to preserve the dataset’s privacy (Cases 01x in Table 1).

**Client (ts)** is where the testing data, which is probably sensitive and confidential, resides and needs analytics. The owner of this data can opt to encrypt it, and this encryption can be provided via simple software components installed on the *client* side available via communication with the *server*. Next, the encrypted testing data is sent to the *server* for prediction. Finally, the encrypted predicted scores are received back. The *client* should be responsible for setting the threshold on the scores for the final hard decision or classification. This is to achieve a required level of aggressiveness to control the per-class sensitivity, such as in the case of the binary classification problem, in which the threshold provides the trade-off between the sensitivity and specificity and thus controls the operating point on the receiver operating characteristic (ROC) curve.

**Server** is the cloud engine for prediction. On the one hand, it interfaces with the *client* to receive the encrypted dataset for prediction and sends back encrypted predictions, and on the other hand, it interfaces with *ML construction* to receive a particular predictive model. Based on the underlying encryption scenario (Table 1), the *server* receives the appropriate public key from these two modules. In addition, for the  $C$ -class classification problem and for greater privacy preservation for the model and/or the dataset (**tr**), the server can optionally multiply the scores  $s_c(x)$ ,  $c = 1, \dots, C$ , where  $x \in \mathbf{ts}$ , by a random integer. This keeps the relative  $C$  scores unaffected. However, this disallows the *client* from inferring information about the model weights ( $\mathcal{M}$ ) by sending pseudo-feature vectors in the form  $x = (0, \dots, 1, 0, \dots)$  (only one feature is 1; the others are zeros).

To illustrate the utility of this simple architecture, we demonstrate how scenario 1 can be implemented in a very practical setup. When ML training is based on public data (**tr**), the weights of the trained model  $\mathcal{M}$  are deployed on the *server* in unencrypted form, while the queries (**ts**) must be encrypted for security sensitivity. Under a hospital’s public key, many parties may also be eligible to upload data (e.g., doctors and patients). The *server* is used for deploying ML implementations  $\mathcal{M}$ . In this case, the hospital sends encrypted data to the *server*. In the *server*, many computations can be done on the encrypted data and the results sent back to the hospital. Only the hospital can decrypt the data because the private key is provided only on the hospital side. In the follow-

ing section, we conduct a large set of ML experiments under the MLE framework and scenario 1 on one of the most recent high-quality genomic datasets.

## 4. Experiments

In this section, we concisely describe the dataset (more details are provided in the supplementary materials), explain the different ML experiments to build the predictive model, demonstrate computational aspects of the encryption process, and finally introduce the open-source platform of the whole project.

### 4.1. MSK-IMPACT Dataset

MSK-IMPACT, a clinical sequencing cohort dataset (Zehir et al., 2017), comprises genomic patient records extracted from tumor-tissue samples taken from 10,336 patients. Since tumors are usually the results of many mutations, there are more than 100,000 discovered mutations. The dataset consists of 11 files linked together with `sample_ID` and `Patient_ID` and contains various information about the somatic mutations mutations within the genomic samples, including mutation signature, copy number alternation, and gene fusion data files. “With maturing clinical annotation of treatment response and disease-specific outcome”, according to (Zehir et al., 2017), “this dataset will prove a transformative resource for identifying novel biomarkers to inform prognosis and predict response and resistance to therapy...Tumor molecular profiling is a fundamental component of precision oncology, enabling the identification of genomic alterations in genes and pathways that can be targeted therapeutically”.

The authors of the dataset tried to associate “biomarkers” with a particular type of cancer using simple methods of association, such as relative frequency. Then, to illustrate the usefulness of their DNA-sequence approach, they leveraged the Oncology Knowledge Base (Chakravarty et al., 2017) to see how many of the mutations they detected (stratified by cancer type) were known to be actionable, that is, have an associated treatment or gene therapy. Recently, a subset of the MSK-IMPACT-2017 dataset with a portion of these features, somewhat lightly engineered, was used by Penson et al. (2020) with a more ML-oriented approach. Using logistic regression (LR), they achieved an overall accuracy of 75% in detecting the cancer type from genetic information; yet, their approach did not consider privacy preservation.

### 4.2. Building the Model

In addition to the predictive power required for any ML model, the objective of privacy preservation via FHE requires the final ML model be FHE-friendly, that is, based only on addition and multiplication operations, as was explained in Section 2. Some ML models cannot satisfy both of these objectives. For example, a random forest (RF) has binary decision splits that are not FHE-friendly. However, although linear models (LM), logistic regression (LR), support vector machines (SVM), and many others are all HE-friendly, they may not perform well on a particular dataset.

For our MSK dataset, we tried 2,240 different ML configurations that were the cross products of five methods for transforming features, two methods for dropping the least important features, four values for the number of dropped features  $p$ , and three classifiers, LR, SVM, RF, each with several sets of tuning parameters. The detailed parameters of these 2,240 experiments are listed in Table 2. The best three configurations of the 2,240 are listed in Table 3 and were achieved by LR, SVM, and RF, respectively. LR, after feature standardization,  $\chi^2$  selection, and dropping the least informative 2,500 features, achieved the highest accuracy of 77.47% (using 10-fold cross validation), which was higher than that obtained by Penson et al. (2020) on the same dataset.

Although the model was trained on MSK, a public dataset that requires no encryption, the final model parameters (Eq. (2)) must be encrypted since they will be multiplied by the encrypted features of the testing data. The next section explains the encryption of both the testing dataset and the parameters of the final trained model. Since we are using FHE, we must convert all floats to integers. To do this, we scale all floats by some  $10^d$ , where  $d$  is the number of decimal places included in the scaled floats, therefore controlling the computational precision. After scaling, we round off any remaining decimal to achieve the final integers. The effect of precision on the accuracy of the best model is illustrated in Table 4, where it is clear that a multiplier of  $10^4$  would be adequate.

### 4.3. Encrypting the Testing Dataset

#### 4.3.1. The SEAL library

Different libraries exist for implementing HE (Carey, 2020); among them, SEAL (Microsoft, 2020) is an open-source HE library developed by the cryptography and privacy research group at Microsoft. The library is written in C++ and can run in many environments. SEAL allows addition and multiplication to be performed on numbers. Other operations, such as encrypted comparison, sorting, and regular expressions, are in most cases not feasible on encrypted data using this technology. SEAL supports two FHE schemes: the Brakerski/Fan-Vercauteren (BFV) scheme, which allows modular arithmetic to be performed on encrypted integers, and the Cheon-Kim-Kim-Song (CKKS) scheme, which allows addition and multiplication on encrypted real or complex numbers, but this latter scheme yields only approximate results.

#### 4.3.2. Computational Aspects of Encryption Operations

From the previous section, a  $C$ -class LR model was the winner for this dataset; formally, this model is given by

$$\Pr(G = c|X = x) = \frac{\exp(s_c)}{1 + \sum_{l=1}^{C-1} \exp(s_l)}, \quad (2a)$$

$$s_c = w_{c0} + w'_c x, \quad c = 1, \dots, C, \quad (2b)$$

where,  $C = 22$  types of cancer, the patient feature vector is  $x \in \mathbf{R}^p$ ,  $p = 5,599$ ; and the testing dataset  $\mathbf{ts}$  to be encrypted has  $N = 7,791$  patient records. By construction, the MLE framework requires sending the encrypted score of each testing observation to the client rather than the final hard decision for

Table 2: Different configurations of feature pre-processing and classifiers tried on the dataset:  $5 \times 2 \times 4 \times (8 + 36 + 12) = 2,240$  different configurations.

Feature Preprocessing			Classifier	Parameters
Transformation	Selection	Reduced $p$		
None		0	LR	penalty: l1, l2; C: 0.1, 1, 10; solver: liblinear, lbfgs
Standardize	MI	2500	SVM	kernel: linear, poly, rbf; penalty: l1, l2; C: 0.1, 1, 10; loss: hinge, squared_hinge
MinMax	$\chi^2$	3500		
$\log(x)$		4500	RF	criterion: gini, entropy; n_estimators: 100, 200, 300; bootstrap: False, True
$\log(x + 1)$				

Table 3: The best three configurations in Table 2. The first two are the LR and SVM, which are HE-friendly.

Feature Preprocessing			Classifier	Parameters	Accuracy (%)
Transformation	Selection	Reduced $p$			
Standardize	$\chi^2$	2500	LR	C: 1; Penalty: l2; Solver: liblinear	77.47
None	$\chi^2$	2500	SVM	kernel: linear; C:0.1; penalty: l2; loss: squared_hinge	77.23
Standardize	MI	2500	RF	criterion: gini; n_estimators: 200; bootstrap: False	73.92

Table 4: Accuracy and percentage of score ranking of the best model at different multiplier precision.

Precision	$10^1$	$10^2$	$10^3$	$10^4$	$10^5$	$10^6$	$10^7$	$10^8$	$10^9$
Accuracy (%)	73.88	77.41	77.47	77.41	77.41	77.41	77.41	77.41	77.41
agreeing ranks (%)	36.49	70.80	95.82	99.41	99.94	100.0	100.0	100.0	100.0

Table 5: Effect of encryption parameters on encryption time. All libraries support automatic selection of `CoeffModulus`. NA indicates noise  $\sim 0$

#	polyModulusDegree	PlainModulus	Time in Sec.
1	8192	2048	3456
2	2048	1024	96
3	1024	512	NA
4	1024	1024	NA
5	2048	1024	87
6	2048	512	88
7	2048	1499	94
8	2048	786433	89

trading off the types of error. In addition, from (2), the numerator is a monotonic exponential function and the denominator is only for scaling, so probabilities sum to 1. Therefore, it is sufficient to encrypt the linear term  $s_c$  and treat it as the final score sent to the *client*. We applied the BFV scheme implementation of SEAL to perform this weighted summation term. The encryption operations are explained as follows. (1) Scale and encrypt the feature list. (2) Multiply encrypted features by plain text weights and sum encrypted values along with the scaled bias step. (3) Decrypt the results and repeat step 2 for each set of coefficients, i.e. each class. We tested different encryption parameters to compare computational time. Table 5 illustrates the computational time required as a function of a subset of the parameter space. Rows 3 and 4, caused the ciphertext noise budget to reach zero. This noise budget is determined by the encryption parameters, and once the noise budget of a ciphertext reaches

zero, it becomes too corrupted to be decrypted. Thus, it is essential to choose parameters large enough to support the desired computations; otherwise, the correct result is impossible to obtain, even with the secret key. The values in row five give the best average per sample prediction time after testing on the entire dataset (7791 records), which spanned over seven days of computations on an i7core-2.5GHz-16G machine. From Eq. (2b), this time is obviously  $T = NC((p + 1)(E + M + A) + D)$ , where  $E$ ,  $M$ ,  $A$ , and  $D$  are the encryption, multiplication, addition, and decryption times, respectively. During this experiment, `IntegerEncoder` was used to encode integers to BFV plain-text polynomials. `IntegerEncoder` is easy to understand and uses simple computations; however, there are more efficient approaches such as `BatchEncoder`, which can be investigated in future works.

#### 4.4. MLE Framework: Opensource and Deployment

The official repository of this project (Briguglio et al., 2021) contains three sets of open-source resources. (1) The Python code that produced all the ML experiments from Section 4 is offered, organized, and commented on to challenge the ML community to develop more accurate, predictive models. (2) The system design and client-server implementation and implementation code of the MLE framework from Section 3 is offered to the software engineering community to propose more system functionalities. (3) A free cloud service that encapsulates the best ML model and the client-server design is offered as a simple end-user interface to individuals in the medical field to test on particular cases. We hope this kind of dissemination to different communities helps the evolution of privacy-preserving ML for precision medicine.

## 5. Conclusion and Suggested Future Work

Toward building privacy preserving Machine Learning (ML) models for precision medicine, this article has three contri-

butions. First, we proposed and implemented a machine learning with encryption (MLE) framework that accommodates different scenarios for encrypting the ML training-testing process. Second, and most importantly, we analyzed the recent high-quality clinical sequencing cohort dataset MSK-IMPACT and provided a predictive model that is both secure and outperforming the most recent predictive model built for the same dataset. Third, we offered the ML, software engineering, and precision medicine communities free resources: respectively, the client-server implementation of the framework, the Python code of all the ML experiments, and a cloud service to test genomic cases. These offerings contribute to the evolution of the privacy-preserving analytics of precision medicine.

From Table 1, it seems that scenario 2 is an open venue for incorporating more accurate ML models on public datasets. As discussed in Section 2, Bost et al. (2015) implemented a very sophisticated scenario that can be classified indirectly as scenario 2. They assumed that the owner of the dataset  $\mathbf{ts}$  is responsible for testing the model  $\mathcal{M}$ . They designed their model encryption in a nested way to preserve its privacy and protect it from the owner of  $\mathbf{ts}$  to learn anything about its structure. Because of our MLE framework, we do not need these constraints, since the computations run on the *server* side, not the *client* side. Therefore, under the MLE framework, future work can benefit from scenario 2 by designing more ML models on datasets and encrypting the model parameters if needed.

## 6. Acknowledgment

The authors thank the National Research Council (NRC) for funding the project under the Collaborative Research and Development Grant (award number CDTs-102).

## References

- Acar, A., Aksu, H., Uluagac, A.S., Conti, M., 2018. A survey on homomorphic encryption schemes: Theory and implementation. *ACM Comput. Surv.* 51, 1–35.
- Aslett, L.J.M., Esperança, P.M., Holmes, C.C., 2015a. Encrypted statistical machine learning: new privacy preserving methods. *arXiv:1508.06845v1*.
- Aslett, L.J.M., Esperança, P.M., Holmes, C.C., 2015b. A review of homomorphic encryption and software tools for encrypted statistical machine learning.
- Bost, R., Popa, R.A., Tu, S., Goldwasser, S., 2015. Machine Learning Classification over Encrypted Data.
- Briguglio, W., Moghaddam, P., Yousef, W.A., Traore, I., Mammun, M., 2021. Machine Learning via Encryption (MLE) Framework in Precision Medicine to Preserve Privacy. <https://github.com/isotlaboratory/Healthcare-Security-Analysis-MLE>.
- Carey, A., 2020. On the Explanation and Implementation of Three Open-Source Fully Homomorphic Encryption Libraries Fully Homomorphic Encryption Libraries.
- Chakravarty, D., Gao, J., Phillips, S., Kundra, R., Zhang, H., Wang, J., Rudolph, J.E., Yeager, R., Soumerai, T., Nissan, M.H., Chang, M.T., Chandralapaty, S., Traina, T.A., Paik, P.K., Ho, A.L., Hantash, F.M., Grupe, A., Baxi, S.S., Callahan, M.K., Snyder, A., Chi, P., Danila, D.C., Gounder, M., Harding, J.J., Hellmann, M.D., Iyer, G., Janjigian, Y.Y., Kaley, T., Levine, D.A., Lowery, M., Omuro, A., Postow, M.A., Rathkopf, D., Shoushtari, A.N., Shukla, N., Voss, M.H., Paraiso, E., Zehir, A., Berger, M.F., Taylor, B.S., Saltz, L.B., Riely, G.J., Ladanyi, M., Hyman, D.M., Baselga, J., Sabbatini, P., Solit, D.B., Schultz, N., 2017. Oncokb: A precision oncology knowledge base. *JCO Precision Oncology*, 1–16.
- Dowlin, N., Gilad-Bachrach, R., Laine, K., Lauter, K., Naehrig, M., Wernsing, J., 2016. Cryptonets: Applying neural networks to encrypted data with high throughput and accuracy, in: 33rd Int. Conf. Mach. Learn. ICML 2016, pp. 342–351.
- Gentry, C., 2009. Fully homomorphic encryption scheme.
- Graepel, T., Lauter, K., Naehrig, M., 2013a. ML confidential : machine learning on encrypted data, in: Kwon, T., Lee, M.K., Kwon, D. (Eds.), *Information Security and Cryptology - ICISC 2012 (15th International Conference, Seoul, Korea, November 28-30, 2012, Revised Selected Papers)*, Springer, Germany. pp. 1–21. Conference; 15th International Conference on Information Security and Cryptography; 2012-11-28; 2012-11-30 ; Conference date: 28-11-2012 Through 30-11-2012.
- Graepel, T., Lauter, K., Naehrig, M., 2013b. ML confidential: Machine learning on encrypted data, in: *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, pp. 1–21.
- Hesamifard, E., Takabi, H., Ghasemi, M., 2017. CryptoDL: Deep Neural Networks over Encrypted Data .
- Microsoft, 2020. Microsoft SEAL (release 3.6). <https://github.com/Microsoft/SEAL>. Microsoft Research, Redmond, WA.
- Nandakumar, K., 2019. Towards Deep Neural Network Training on Encrypted Data. *IEEE Conf. Comput. Vis. Pattern Recognit. Work.*
- Penson, A., Camacho, N., Zheng, Y., Varghese, A.M., Al-Ahmadie, H., Razavi, P., Chandralapaty, S., Vallejo, C.E., Vakiani, E., Gilewski, T., et al., 2020. Development of genome-derived tumor type prediction to inform clinical cancer care. *JAMA oncology* 6, 84–91.
- Rahman, M.S., Khalil, I., Alabdulatif, A., Yi, X., 2019. Privacy preserving service selection using fully homomorphic encryption scheme on untrusted cloud service platform. *Knowledge-Based Syst.* 180, 104–115.
- Rivest, R., Adleman, L., Dertouzos, M., 1978. On Data Banks And Privacy Homomorphism. Technical Report. Massachusetts Institute of Technology.
- Sathya, S.S., Vepakomma, P., Raskar, R., Ramachandra, R., Bhattacharya, S., 2018. A Review of Homomorphic Encryption Libraries for Secure Computation .
- Zehir, A., Benayed, R., Shah, R.H., Syed, A., Middha, S., Kim, H.R., Srinivasan, P., Gao, J., Chakravarty, D., Devlin, S.M., et al., 2017. Mutational landscape of metastatic cancer revealed from prospective clinical sequencing of 10,000 patients. *Nature medicine* 23, 703.

## Supplementary Material

### 7. Genomics

In this section we provide a more detailed description for the MSK-IMPACT dataset (Zehir et al., 2017).

#### 7.1. Synopsis: MSK-IMPACT 2017

##### *Data\_CNA:*

In this file, each column is a sample, and each row is a feature. Each feature is represented by a string (e.g., EGFR) which is the HUGO symbol for the corresponding gene. The feature values are floats, which represent the copy number alteration of the feature's gene. Positive and negative numbers represent duplication or deletion of repeat nucleotides respectively.

##### *Data\_Fusion:*

This file structure is not as straightforward from an ML perspective. Some general information is easily parsable, but more specific information contained in the comments column requires more creative approaches and possibly expert (biochemical background) consultation. The file comprises many rows, each with a gene HUGO symbol and sample ID, among other fields (see below). Each row describes half a fusion in which two genes are mixed together via deletion, inversion, or translocation. Since two genes are involved, two rows describe each half of the fusion effect. The only time this is not true is when the fusion is intergenic. This is when a gene is mixed with DNA material which is located between genes, that is, is non-coding. In this case only the information for the coding DNA material in the fusions is listed in a single row. The remaining fields in each row are as follows:

- **Huge\_Symbol:** unique gene identified for the 410 genes sequenced with some being post-fixed with an Arabic numeral, a Latin letter, or in special cases both to indicate membership in a gene family
- **Entrez\_Gene\_ID:** a gene integer ID used by the National Center for Biotechnology Information (NCBI), only present for two samples
- **Center:** where the sample was taken; set to "MSKCC-DMP", signifying the Memorial Sloan Kettering Cancer Center, for all samples in the file
- **Tumor\_Sample\_Barcode:** sample ID
- **Fusion:** indicating which gene(s) are involved in the fusion
- **DNA support:** "yes" for all samples, indicating that fusion was detected with DNA sequencing
- **RNA support:** "unknown" for all samples, indicating that RNA-sequencing was not preformed
- **Method:** is "NA" for all samples
- **Frame:** either "unknown", "out-of-frame", or "in-frame" indicating the effect of each fusion
- **Comments:** short comments written by practitioners giving specifics of the mutation

##### *Data\_SV:*

structural variation, is a more general classification for mutations than fusion mutations or copy number variations. There are 31 categorical, numerical, and text features describing the type, location, and prevalence of the structural variances.

- **Annotation:** specific type and location of structural variation
- **Breakpoint\_Type:** the type of breakpoint, that is, the junction between normal and rearranged
- **Comments:** small notes on some mutations
- **Confidence Class:** indicates the confidence in the final sequencing, and whether it was automatically or manually determined

The remaining features are not documented in a single source but detail various aspects of the structural variations such as sequencing method, the quality of sequencing, comparison with germline DNA, variation location, and so on.

##### *Data\_Mutation\_Significance\_Contribution:*

This file contains 30 numerical features corresponding to known mutation signatures. For each sample, the feature value is the percentage of mutations explained by the corresponding signature

##### *Data\_Mutation\_Significance\_Confidence:*

30 numerical features that correspond to the same mutation signature found above, but the feature value is the confidence in the contribution scores instead. The mechanism used to determine confidence in contribution score is described in Huang (2018)<sup>1</sup>

##### *Data\_Mutation\_mskcc:*

This file contains 46 columns describing mutations using a subset of the columns found in the mutation annotation format (MAF) format. All columns and their descriptions can be found in the GDC MAF Format<sup>2</sup>, except two:

- **Hotspot:** zero for all samples; was not configured when the table was made, so the same label was applied to all mutations
- **cDNA\_change:** the nucleotide change described with HGVS expression

##### *Data\_Mutation\_extended:*

This is identical to Data\_Mutation\_mskcc but is missing the "cDNA\_change" column

##### *Data\_Gene\_Panel\_Matrix:*

This file records which type of panel was used to extract genomic data from the sample: the 341 gene or 410 gene panel. The 341 genes are a subset of the 410 genes, so this distinction does not manifest in the other records because when creating a record from a sample that used the 341 gene panel; the 69 genes are not tested for were marked as not detected

<sup>1</sup><https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5860213/>

<sup>2</sup>[https://docs.gdc.cancer.gov/Data/File\\_Formats/MAF\\_Format/](https://docs.gdc.cancer.gov/Data/File_Formats/MAF_Format/)

#### *Data\_cna\_hg19:*

This file is the output of [DNACopy](#), which is an open-source software package written in R that “identifies genomic regions with abnormal copy number”, that is, copy number alterations. Each row in the file corresponds to one such region or “segment” and has five columns describing it:

- ID: the sample ID
- chrom: the [chromosome](#) the segment is in
- loc.start: start location of the segment
- loc.end: end location of the segment
- num.mark: number of probes bound to the segment
- seg.mean: the mean value, across all probes, of the segment; represents the  $\log_2$ -ratio of tumor copy number to normal copy number; a positive value indicates that the tumor has a higher copy number and vice versa

#### *Data\_clinical\_patient:*

Each row corresponds to a patient and contains five columns:

- #Patient Identifier: the patient ID
- Sex: the patient’s gender
- Patient’s Vital Status: whether they are deceased or alive
- Smoking History: whether they previously, currently, or never smoked
- Overall Survival (Months): How long they survived since initial diagnosis; blank if they are currently alive or died after last follow up
- Overall Survival Status: the same as Patient Vital Status with a slightly different format

#### *Data\_clinical\_sample:*

Each row corresponds to a sample and contains 16 columns:

- #Patient Identifier: the patient ID
- Sample Identifier: the sample ID
- Sample Collection Source: whether the sample was collected in house at MSKCC or by another party
- Specimen Preservation Type: method used to preserve the sample
- Specimen Type: how the specimen was collected
- DNA Input: amount of DNA in the sample in nanograms
- Sample Coverage: number of unique reads that included a given nucleotide during sequencing
- Tumor purity: percentage of cancer cells in sample
- Matched Status: whether the patient was matched with a gene therapy
- Sample Type: whether the sample came from the primary or metastatic tumor
- [Primary site](#): the location of the primary tumor
- Metastatic site: where the [metastasis](#) occurred

- Sample Class: type of cancer tissue (is tumor for all samples)
- OncoTree Code: [unique code](#) for specific tumor type
- Cancer Type: type of cancer that caused the tumor
- Cancer Type: Detailed: more specific sub-type of the cancer

#### 7.2. Glossary

**breakpoint** the beginning or end point of a structural variation

**chromosomal inversion** mutation in which a segment of a chromosome is reversed

**chromosomal translocation** two types; Robertsonian translocation, which occurs when two [non-homologous chromosomes](#) get attached, and reciprocal translocation, which occurs when parts are exchanged between two non-homologous chromosomes

**chromosome** a molecule that contains part of the [genome](#) in a condensed, manageable package

**copy number alteration** a change in copy number that has arisen in any cell of the body after conception

**fusion** when two previously independent genes are combined or fused together that results from [chromosomal translocation](#), [interstitial deletion](#), and [chromosomal inversion](#)

**gene family** a set of several similar genes formed by duplication, generally with similar function

**genome** genetic material of an organism

**HUGO symbol** unique gene identifier derived from the [HUGO gene nomenclature](#)

**in-frame** a mutation, where the translation into protein is not completely disrupted, creating still-functional proteins

**interstitial deletion** a mutation in which part of the DNA (not including the terminal portion of a chromosome) is left out during DNA replication

**metastasis** secondary malignant growth distant from the primary site of cancer

**mutation** alteration of the nucleotide sequence of the genome

**mutation signatures** characteristic combinations of mutations arising from specific mutation processes

**out-of-frame** a mutation, where the translation into a protein is completely disrupted, creating non-functional proteins.

**primary site** the location on the body where the first tumor progression begins

**somatic mutation** genetic alteration acquired by cells that are the progeny of cancerous cells

**structural variation** any kind of structural variation to the genome

## 8. Homomorphic Encryption

Below is an example presented by [Sathya et al. \(2018\)](#) to introduce the high-level concept of Homomorphic Encryption (HE).

1. Let  $m$  be the plain text message.
2. Let a shared public key be a random odd integer  $p$ .
3. Choose a random large  $q$ , small  $r$ ,  $|r| \leq p \div 2$ .
4. Ciphertext  $c = pq + 2r + m$  (ciphertext  $c$  is close to multiple of  $p$ ).
5. Perform homomorphic addition/multiplication as required.
6. Decrypt:  $m = (c \bmod p) \bmod 2$ .

Homomorphic addition can be illustrated as follows

$$c_1 = q_1 \times p + 2 \times r_1 + m_1 \quad (3a)$$

$$c_2 = q_2 \times p + 2 \times r_2 + m_2 \quad (3b)$$

$$c_1 + c_2 = (q_1 + q_2) \times p + 2 \times (r_1 + r_2) + (m_1 + m_2), \quad (3c)$$

and Homomorphic multiplication as follows

$$c_1 = q_1 \times p + 2 \times r_1 + m_1 \quad (4a)$$

$$c_2 = q_2 \times p + 2 \times r_2 + m_2 \quad (4b)$$

$$c_1 \times c_2 = ((c_1 \times q_2) + q_1 \times c_2 \times q_1 \times q_2) \times p + 2(2 \times r_1 \times r_2 + r_1 \times m_2 + m_1 \times r_2) + m_1 \times m_2. \quad (4c)$$

Although homomorphic encryption holds massive potential in theory, it suffers from notable shortcomings in practice. In many cases, it is limited to only addition and multiplication meaning many functions must be approximated with high degree polynomials which incur a large computational overhead. Even when using only this subset of operations, homomorphic operations are orders of magnitude slower than conventional operations on plaintext data. Homomorphic encryption also leads to substantial ciphertext expansion of a magnitude proportional to the targeted security strength. Further, homomorphic encryption schemes do not allow unlimited operations without first decrypting and re-encrypting or running an expensive denoising operation. These and other considerations make homomorphic encryption not easily adaptable to practical applications without substantial foresight and planning. For a more in depth review of the limitations and practical considerations of homomorphic encryption ([Aslett et al., 2015b](#)).