Article

Data Obfuscation for Privacy-Preserving Machine Learning using Quantum Symmetry Properties

Sebastian Raubitzek * , Alexander Schatten , Sebastian Schrittwieser , Kevin Mallinger

Posted Date: 23 July 2024

doi: 10.20944/preprints202407.1701.v1

Keywords: Data Obfuscation; Machine Learning; Boost Classifier; Medical Data; Diabetes; Breast Cancer; Artificial Intelligence; Data Privacy; Quantum Machine Learning; Quantum Information Processing

Article

Data Obfuscation for Privacy-Preserving Machine Learning using Quantum Symmetry Properties

Sebastian Raubitzek [1,*], Alexander Schatten [1], Sebastian Schrittwieser [2] and Kevin Mallinger [1]

[1] aSBA Research gGmbH, Floragasse 7/5.OG, Vienna, 1040, Austria
[2] bChristian Doppler Laboratory for Assurance and Transparency in Software Protection, Research Group Security & Privacy, Faculty of Computer Science, University of Vienna, Kolingasse 14-16, Vienna, 1040, Austria
* Correspondence: sraubitzek2@sba-research.org

Abstract: This study introduces a data obfuscation technique, leveraging the exponential map associated with the generators of Lie groups. Originating from quantum machine learning frameworks, our method illustrates the practical application of quantum mechanics principles in data processing. Specifically, it employs the exponential map of a generator algebra to introduce controlled noise into the data, achieving obfuscated data while preserving its utility for machine learning tasks. This strategy is shown to safeguard privacy in sensitive datasets, such as discussed medical records, and to enhance dataset volume and diversity through augmentation. Our empirical analysis, benchmarked against standard machine learning approaches, demonstrates that our method can maintain or even improve the predictive accuracy of the original data. This research highlights the potential of Lie group theory for advancing data privacy in medicine, marking a significant contribution to machine learning methodologies by offering the dual benefits of data obfuscation and enrichment. Through this synthesis of algebraic structures and machine learning, we propose new pathways for the secure and effective use of data in sensitive areas.

Keywords: data obfuscation; machine learning; boost classifier; medical data; diabetes; breast cancer; artificial intelligence; data privacy; quantum machine learning; quantum information processing

1. Introduction

Quantum technologies are increasingly being integrated into various disciplines, ranging from Quantum Key Distribution (QKD) [85], which enables secure key exchanges, to the advancement of quantum information processing technologies like quantum computers [87]. Furthermore, Quantum Machine Learning (QML) emerges as a promising field, leveraging quantum computational advantages to address complex problems [79]. Despite the fact that the physical realization of quantum computers and quantum circuits currently trails behind their classical counterparts, the theoretical and conceptual frameworks of quantum technologies have demonstrated promising potential across a broad spectrum of applications. Here, QML represents a frontier in computational science, blending quantum computing's potential with classical machine learning's algorithmic precision. The promise of QML lies in its capacity to process and analyze complex, high-dimensional data sets beyond the reach of current classical methodologies [79]. Central to these quantum information technologies are feature maps, which are instrumental in encoding classical data onto quantum circuits or qubits. These processes rely on the algebraic principles underlying the symmetries of the SU(2) Lie group [19], and exactly these underlying algebraic structures of information encodings are the focus of this article.

Lie groups, which are continuous transformations, are instrumental in numerous physical and mathematical theories, providing a rich lexicon for describing symmetries and corresponding transformations [9,93].

This paper introduces a novel approach to data obfuscation using the exponential map of Lie group generators, tested for publicly available medical data. Similar to how quantum computing aims to harness the multidimensional and symmetrical qualities of quantum states, our method applies these same properties, specifically the symmetries found in Lie groups, to alter data within a high-dimensional space. This connection highlights our methodology's interdisciplinary approach, effectively merging concepts from quantum mechanics, algebra, and machine learning in a novel way.

Central to our investigation is the question: *Can Lie Group theory effectively obfuscate sensitive data while retaining its essential properties for successful machine learning applications, thereby preserving the accuracy and informational value of the original dataset?*

Our research builds on the foundational work of Schuld et al. [11,31,44] and IBM's Qiskit [12] in the field of Quantum Machine Learning (QML) and their exploration of feature maps for projecting data onto qubits.

Our main contributions in this article can be summarized as:

- We develop a novel data obfuscation framework using the exponential map of Lie group generators, tailored for privacy-preserving processing of medical data used in machine learning approaches.
- We show where and how the invertibility of our obfuscation technique breaks down by injecting noise into the exponential map of Lie group generators. Thus making it impossible to recover the original data.
- We demonstrate the efficacy of this approach in maintaining and occasionally surpassing the predictive accuracy of machine learning models compared to non-obfuscated datasets.
- We establish a conceptual link between the principles of quantum machine learning and our obfuscation methodology, highlighting the potential for cross-disciplinary innovation in leveraging symmetries for data privacy, thus showing the applicability of quantum mechanical concepts in this context.

The remainder of this article is organized as follows: We provide a collection of related related work in Section 2. Section 3 provides our methodology, i.e., a background on quantum feature maps, Lie groups, how to use them for data obfuscation, and where invertibility of the exponential map breaks down by injecting noise. Section 4 describes our experimental setup and the employed data sets (Section 4.1). The following Section 4.2 presents our results. The final Section 5 concludes our approach and findings, discusses the implications, and gives an outlook on future applications.

2. Related Work

The protection of patient privacy is paramount in medical data processing, making data obfuscation a critical area of research. Data obfuscation techniques aim to mask sensitive information while maintaining the utility of the data for machine learning applications. This ongoing research is vital as it addresses the dual challenge of protecting patient confidentiality and enabling the extraction of actionable insights from medical data [10].

One common approach is data anonymization, where identifiers such as names and social security numbers are removed or replaced with pseudonyms. For instance, the k-anonymity model [108] ensures that each record is indistinguishable from at least k-1 others regarding certain attributes. However, [109] highlighted the vulnerability of k-anonymity to re-identification attacks, leading to the development of more sophisticated methods. Studies by Lu et al. [111] have applied homomorphic encryption to medical datasets, enabling secure analysis without compromising patient confidentiality.

Deep learning (DL)-based algorithms for image classification have demonstrated remarkable results in improving healthcare applications' performance and efficiency. To address privacy concerns, especially in cloud-based solutions, data obfuscation techniques like variational autoencoders (VAEs) combined with random pixel intensity mapping can be used for enabling DL model training on secured medical images while ensuring privacy [107].

Olatunji et al.'s comprehensive review [10] of healthcare data anonymization techniques underscores the delicate balance between privacy and utility in the context of modern big data and machine learning challenges, which also applies to data obfuscation.

Quantum information processing presents novel opportunities for advancing machine learning, particularly through quantum machine learning (QML). The integration of quantum computing with machine learning algorithms has the potential to revolutionize data processing, offering significant improvements in speed and efficiency.

A key concept in QML is the use of quantum feature maps, which embed classical data into high-dimensional quantum states. This process can enhance the representational capacity of machine learning models. Havlíček et al. [13] demonstrated that quantum feature maps could enable the classification of complex datasets that are challenging for classical models.

Quantum algorithms can potentially provide new methods for data privacy. For example, Lloyd et al. [112] proposed quantum algorithms for principal component analysis (PCA), which can be applied to obfuscate data while preserving essential features for machine learning tasks. Such approaches leverage the principles of quantum mechanics to enhance data security and utility simultaneously.

Synthetic data generation [113] involves creating artificial datasets that resemble real data but do not contain actual patient information. Techniques such as generative adversarial networks (GANs) have been employed to generate realistic medical data for machine learning. While synthetic data can effectively preserve privacy, ensuring the fidelity and utility of such data is an area of active investigation. For a similar purpose, but to create exemplary test classification datasets, Raubitzek et al. [81] showed that one can use Lie algebras to create synthetic and artificial data. This approach was tested using both quantum machine learning and classical machine learning algorithms.

3. Methodology

We start this section by discussing fundamentals of quantum information processing and quantum machine learning necessary to understand our ideas, which we then expand to present our novel approaches.

Quantum machine learning consists of two steps: First, the feature encoding step, and second the actual quantum computation, whereas we focus solely on the first step.

We describe a standard quantum feature encoding as:

$$|\psi(x)\rangle = U_{\Phi}(x)\,|0\rangle\,, \tag{1}$$

where:

- $|\psi(x)\rangle$ denotes the quantum state obtained by applying the feature map $U_{\Phi}(x)$ to the initial state $|0\rangle$,
- $|\cdot\rangle$ represents a state vector in the complex Hilbert space $\mathcal{H}$,
- $U_{\Phi}(x)$ is a unitary operation encoding classical data x into a quantum state, preserving total probability,
- $|0\rangle$ is the quantum system's initial, "empty" state before encoding.

This equation captures the transformation of classical data x into a quantum state $|\psi(x)\rangle$ through a unitary feature map $U_{\Phi}(x)$.

These feature maps, i.e., unitary transformations, especially those of the Pauli class, are based on SU(2) symmetry properties, meaning that there are matrix transformations that follow certain rules to project arbitrary data on a qubit. The behavior of these Pauli-class feature maps is governed by the Pauli matrices which are three 2 × 2 complex matrices:

$$\sigma_x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \quad \sigma_y = \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \quad \sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}$$

These matrices form a basis for the Lie algebra of the SU(2) group. SU(2) is the group of 2 × 2 unitary matrices with determinant 1, and its Lie algebra, denoted as $\mathfrak{su}(2)$, consists of all 2 × 2 traceless Hermitian matrices.

The Lie algebra $\mathfrak{su}(2)$ is then spanned by the Pauli matrices multiplied by $\frac{1}{2}i$:

$$\mathfrak{su}(2) = \left\{ \frac{1}{2}i\sigma_x, \frac{1}{2}i\sigma_y, \frac{1}{2}i\sigma_z \right\} \tag{2}$$

However, we are expanding this concept to Lie groups SU(n) and SL(n) Lie-groups and add noise to the generators to slightly break these symmetries to obfuscate our original data, thus making the data non-reproducible. To show this, we will first discuss quantum feature maps, and here, two particular ones used in IBM's Qiskit, [12], show where the mechanics of Lie groups are used, and expand this approach to use SU(n) and SL(n), [9,93].

Among various and custom quantum feature maps, the Z and ZZ feature maps are standard choices implemented in IBM'S Qiskit, [12]. These feature maps basically use the properties of Pauli matrices to generate rotations in a complex two-dimensional space to encode classical data into the quantum realm. The basic idea here is that similar to a standard rotation matrix, parameterized using an angle $\theta \in [0, 2\pi]$, one expands this methodology to complex rotations which are parameterized using the Pauli-matrices. This gives rise to the two following feature maps, which are variations of Equation 1, and depicted in Figure 1:

- **The Z Feature Map**

  The Z feature map employs the Pauli-Z operator to encode classical data into quantum states. For a given data point x, it applies a phase rotation to each qubit in a quantum register, proportional to the corresponding feature value in x. Mathematically, this operation is described by:

  $$U_{Z,j}(x) = \exp\left(i x_j Z_j\right), \tag{3}$$

  where Z_j is the Pauli-Z matrix acting on the j-th qubit, and x_j is the j-th component of x. This results in a rotation around the Z-axis of the Bloch sphere, effectively encoding the data within the phase of the quantum state, depicted in Figure 1.
- **The ZZ Feature Map**

  Building on the Z feature map, thus employing the same rotation transformations, the ZZ feature map introduces entanglement between qubits to enrich the feature space. It uses two-qubit gates controlled by the product of pairs of classical data features, depicted in Figure 1

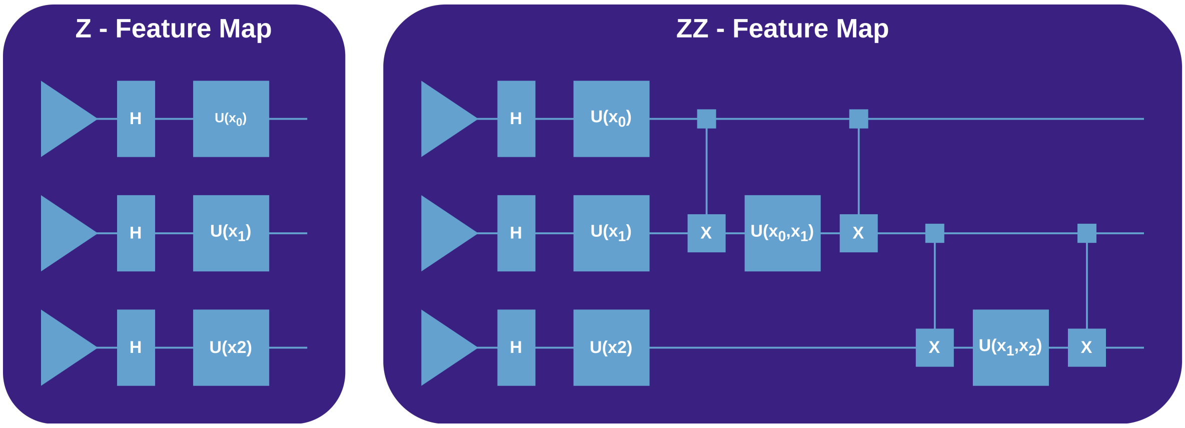

Figure 1. Depiction of the Z- and the ZZ feature maps. Both schemes were adopted according to the implementation in IBM's Quiskit [12]. The triangle on the left symbolizes the incoming qubit; afterward, for both feature maps, the qubit is entangled via a Hadamard gate to make the qubit more expressive. Next, the Pauli-Z-rotations are applied to the qubits.

In quantum information processing, efficient data encoding into a quantum circuit is essential. For each data feature, a corresponding manipulation is required. Our approach expands beyond standard quantum feature encoding, which uses SU(2) transformations for individual qubits. For instance, encoding 8 features requires at least 8 generators. This requirement is met, e.g., by the group SU(3), which has 8 generators, i.e., the Gell-Mann matrices. We parameterize these matrices with normalized features within the exponential map, producing a group element that is applied to a normalized vector, resulting in a complex three-component vector encoding the data sample information. This method permits the application of arbitrary Lie groups for data encoding, assuming the group's generators are constructible. This concept and and a corresponding example are depicted in Figure 2.

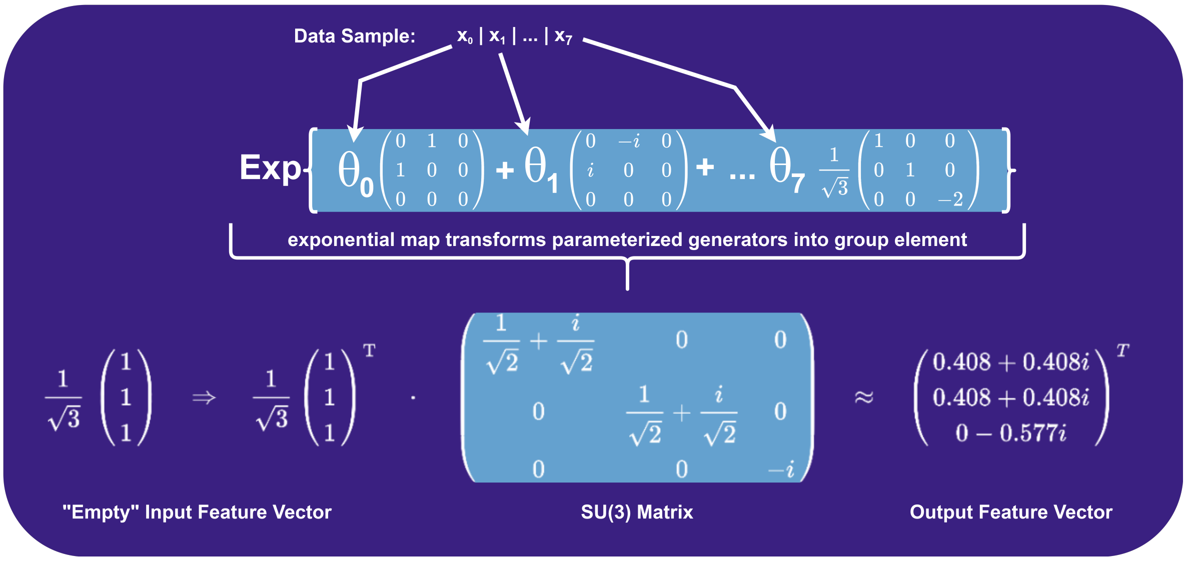
Data Sample:
$x_0 \mid x_1 \mid \dots \mid x_7$
$\text{Exp}\{\theta_0 \begin{pmatrix}0&1&0\\1&0&0\\0&0&0\end{pmatrix} + \theta_1 \begin{pmatrix}0&-i&0\\i&0&0\\0&0&0\end{pmatrix} + \dots\ \theta_7 \frac{1}{\sqrt{3}}\begin{pmatrix}1&0&0\\0&1&0\\0&0&-2\end{pmatrix}\}$
exponential map transforms parameterized generators into group element
$\frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix} \Rightarrow \frac{1}{\sqrt{3}}\begin{pmatrix}1\\1\\1\end{pmatrix}^{T} \cdot \begin{pmatrix}\frac{1}{\sqrt{2}}+\frac{i}{\sqrt{2}}&0&0\\0&\frac{1}{\sqrt{2}}+\frac{i}{\sqrt{2}}&0\\0&0&-i\end{pmatrix} \approx \begin{pmatrix}0.408+0.408i\\0.408+0.408i\\0-0.577i\end{pmatrix}^{T}$
"Empty" Input Feature Vector
SU(3) Matrix
Output Feature Vector

Figure 2. Illustration of our data encoding strategy using the SU(3) group. A data sample with 8 features parameterizes the Gell-Mann matrices, which are then transformed into a group element via the exponential map. This group element is applied to a normalized "empty" input vector, yielding a complex three-component vector that embeds the information of the data sample. Note that imaginary unit *i* is part of *Exp*.

In our approach, we identify a Lie group that is sufficiently large, specifically one within the families of SU(n) or SL(n), which has an adequate number of generators, i.e., more or equal. We use our normalized feature vector $\vec{x} = (x_0, x_1, \dots, x_m)$ to parameterize the generators, thereby obtaining the corresponding group element $U(\vec{x})$:

$$U(\vec{x}) = \exp\left(i \sum_j x_j T_j\right), \tag{4}$$

where x_j are the individual components of the feature vector, T_j are the generators of the selected symmetry group, and $U(\vec{x})$ is a $k \times k$ matrix representing the group element. Should the number of generators exceed the number of features, we set the parameters for the excess generators to zero. This encoding transforms our data samples or vectors into a new feature space and feature vector represented by:

$$\vec{\phi} = \frac{1}{\sqrt{k}} \begin{pmatrix}1\\ \vdots \\ 1\end{pmatrix} \cdot U(\vec{x}) \quad , \tag{5}$$

For the following machine learning process, we also separate real and complex components of the so-obtained feature vector, thus obtaining $2 \times k$ features.

Incorporating a noise term χ to each set of generators guarantees the data to be obfuscated. This is mathematically represented by a random uniform noise component added to each component of the summed up set of parameterized generators, if the parameterized set of generators is complex, we add both a real and a complex noise component. This results in the following expression for our noisy group elements:

$$U_\chi(\vec{x}) = \exp\left(i\sum_j x_j T_j + \chi\right), \tag{6}$$

This addition of noise effectively perturbs each group element $U(\vec{x})$ generated by the exponential map, leading to a slightly altered encoded quantum state.

This expansion of feature maps to arbitrary Lie groups enhances our ability to represent and manipulate data. By leveraging the diverse symmetries and structures of different Lie groups, we can design feature maps that are tailor-made for specific types of data or learning tasks.

Mathematically, adding a small noise vector χ ensures that the perturbed quantum state remains within a vicinity of the original state near the manifold, preserving the relative distances and geometric relationships crucial for machine learning algorithms. This proximity guarantees that while the data is obfuscated enough to protect privacy, it retains sufficient structure for effective learning.

Finally, we can apply the feature map from Equation 6 to each sample several times, every time with a different noise component, and thus use our approach not only to obfuscate data but also to increase the amount of data, i.e., synthesize additional data, thus multiplying the amount of data.

3.1. Retrieving the Original Data

Given the previously outlined discussion on constructing our data obfuscation based on the exponential map of a Lie group, we want to ensure that our original data is not retrievable, which we do using the following construction of our noise component χ.

First of all, we need to make some assumptions about our discussion. We need to assume first that an attacker that wants to acquire the original data is familiar with our obfuscation approach and with Lie groups, corresponding algebras, etc. Then we need to assume an attacker knows about our base vector, as discussed in Equation 5, and finally, the attacker is capable of reproducing the transformation matrix from our transformed feature vector, i.e.

$$\vec{\phi}_\chi = \frac{1}{\sqrt{k}}\begin{pmatrix}1\\ \vdots\\ 1\end{pmatrix} \cdot U_\chi(\vec{x}) \quad , \tag{7}$$

thus reproducing $U_\chi(\vec{x})$. This starts the discussion on how to choose the noise such that one cannot retrieve the original features $\vec{x}$ from our transformation matrix $U_\chi(\vec{x})$.

First, we need to discuss if and when the exponential map of a Lie group is invertible:

Local Invertibility

The exponential map, denoted as $\exp : \mathfrak{g} \to G$, where $\mathfrak{g}$ is the Lie algebra of a Lie group G, is locally invertible around the identity element of G. This follows from the Inverse Function Theorem, which applies because the differential of the exponential map at the identity (zero in the Lie algebra) is the identity map, making it a local diffeomorphism at this point.

Global Invertibility

Globally, the exponential map is generally not invertible. This is because the map can be neither injective (one-to-one) nor surjective (onto):

- **Injectivity:** The exponential map is not injective if there exist elements $X, Y \in \mathfrak{g}$ such that $X \neq Y$ but $\exp(X) = \exp(Y)$. This can occur, for example, when X and Y differ by a multiple of $2\pi i$ in certain directions in $\mathfrak{g}$, particularly for compact or periodic dimensions of G.
- **Surjectivity:** The exponential map may not be surjective for some Lie groups, meaning not all elements of the group can be expressed as the exponential of some element in the algebra. A

typical example is non-connected groups where the exponential map reaches only the connected component of the identity.

Given these arguments, we need to look at the most extreme case: Injectivity and Surjectivity are given globally for a particular Lie group, and the attacker knows which Lie group we used to encode our data and further knows the set of generators we used. Thus we construct our noise in the following way to make our original data non-retrievable:

$$U_\chi(\vec{x}) = \exp\left(I\sum_j x_j T_j + \chi\right) \tag{8}$$

This noise χ can be decomposed into two components:

$$\chi = \chi_G + \zeta \tag{9}$$

where χ_G denotes noise that can be expressed as a linear combination of the generators of the Lie group (with a different parameterization vector $\vec{c}$), and ζ is a residual noise matrix that cannot be expressed as a linear combination of the generators. This results in the following cases: If the matrix $\chi_G \neq 0$ and $\zeta = 0$, then the following occurs. As discussed before, given the most extreme case that one can reconstruct the generators. One can obtain a feature vector from assigning different parameterizations to these generators. However, one cannot retrieve the original feature vector exactly. The features will have a small deviation in each of its components. This means the noise injected into the exponential map slightly distorts the original features. Therefore, we construct χ_G such that:

$$\chi_G = \sum_j \epsilon_j T_j\,, \text{ where } \sum_j |\epsilon_j| = \epsilon\,. \tag{10}$$

Here, ϵ is a controllable parameter, i.e., the level of noise that we inject into our data set. Further, we distribute ϵ randomly among the coefficients $\epsilon_j \neq 0$. Concluding again, one cannot retrieve the original feature vector except one knows precisely the random numbers/coefficients ϵ_j.

The next case we need to discuss is if our residual noise component is not zero, $\zeta \neq 0$, and assuming that $\chi_G = 0$. In this case, the resulting matrix $U_\chi(\vec{x})$ from applying the exponential map, i.e. $U_\chi(\vec{x}) = \exp\left(I\sum_j x_j T_j + \chi\right)$, is not part of the regarded symmetry group, thus our initial symmetry is broken, and we leave the Lie group's manifold. However, this means:

- **Loss of Group Structure:** The resulting matrix is no longer guaranteed to satisfy the properties (closure, associativity, identity, and invertibility) that define the group. Hence, it cannot be inverted within the context of the group.
- **Breaking Symmetry:** The exponential map is no longer mapping elements of the Lie algebra to the Lie group, breaking the symmetry and making the inverse mapping undefined.
- **Non-recoverability of Original Features:** Since the transformation is no longer within the group, one cannot apply the inverse of the exponential map to recover the original features. The noise ζ introduces components that do not belong to the algebra, hence the original structure and information are obfuscated beyond recoverability.

In conclusion, the introduction of residual noise ζ that cannot be expressed as a linear combination of the generators fundamentally disrupts the structure and invertibility of the exponential map, ensuring that the original feature vector cannot be reconstructed from the transformed vector. Further, the noise injected into the parameterizations of the regarded generators ensures a slight distortion of the original features, which further obfuscates the original data. Thus we conclude, the obfuscated data cannot be reconstructed.

doi:10.20944/preprints202407.1701.v1

4. Experiments

We performed experiments on four data sets to measure if the data obfuscated using our augmented noisy Lie group approach can still be classified with a machine learning approach. This means we transform all four data sets with varying amounts of noise and multipliers (i.e., synthetic data), perform a machine learning classification with 80% training data and 20% test or validation data, and note the accuracy of the machine learning prediction on the test data. We also compare this accuracy to the same machine learning approach but without obfuscating/transforming the data. This experimental design is depicted in Figure 3. In the following, we discuss details of our approach, such as the normalization, the employed machine learning algorithms, and the regarded data sets.

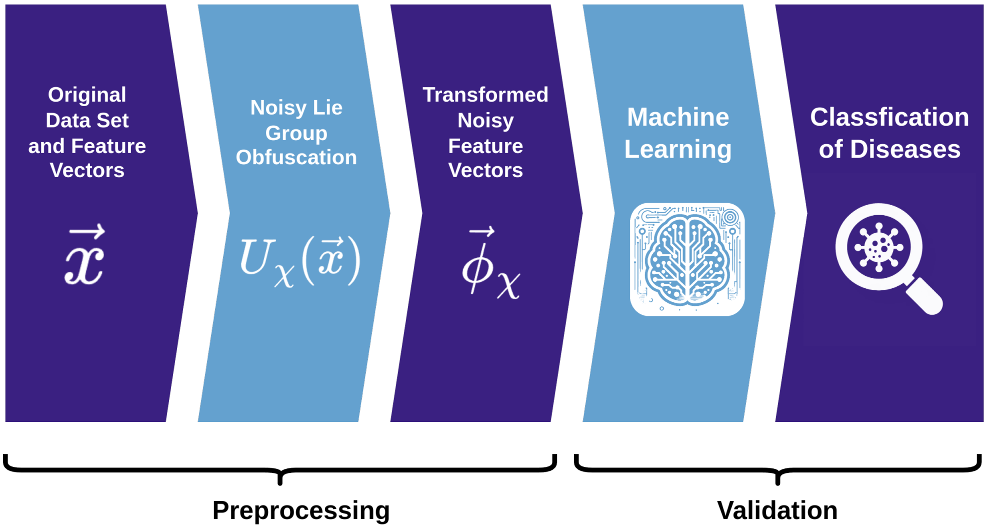
Original Data Set and Feature Vectors
$\vec{x}$
Noisy Lie Group Obfuscation
$U_\chi(\vec{x})$
Transformed Noisy Feature Vectors
$\vec{\phi}_\chi$
Machine Learning
Classfication of Diseases
Preprocessing
Validation

Figure 3. Pipeline of our conducted machine learning experiments. The pipeline depicts the incoming original data as the feature vector $\vec{x}$. Then, the noisy Lie group transformations from Section 3 are applied to the original data set to obtain the transformed feature vector $\vec{\phi}$. This transformed feature vector is then used as the input for the employed machine learning algorithm to classify the individual diseases for each data set.

Normalization of Features: We normalize all features to the range $[0, \pi]$ to effectively utilize the exponential map with our chosen Lie groups. Further, all categorical features were projected into a numerical space such that we give each category a distinct value between 0 and π.

Datasets and Data Augmentation: We employ four distinct datasets, each subjected to five levels of noise and data augmentation, i.e., we change the noise parameter where the noise is sampled from. Data synthetization was performed by multiplying the dataset size by factors ranging from 1 (no augmentation) up to 5, i.e., creating different noisy samples for each data point.

Bayesian Optimization and LGBM Classifier: For the classification tasks, we utilized a Light Gradient Boosting Machine (LGBM) classifier. LGBM is known for its efficiency and effectiveness in handling large datasets and high-dimensional feature spaces, making it an apt choice for our experiments [8]. Bayesian optimization with 100 iterations was employed to search through the hyperparameter space, ensuring the optimal configuration for each experimental condition.

Evaluation Strategy: The datasets were split into training and testing sets using an 80/20 ratio. The performance of the LGBM classifier, trained on the feature-mapped data, was compared against the same LGBM implementation with standard preprocessing, scaling the data to the interval [0,1] for both numerical and categorical features. We chose the standard accuracy score as our primary metric for evaluation.

4.1. Datasets

The following data sets were used to verify if our obfuscation technique can maintain a reasonable performance in classification tasks for medical use cases. All four data sets are medical data sets and are binary classifications, such that the outcome is an identified disease or not. All four data sets are publicly available and can easily be fetched from online databases, e.g., via Python scripts.

1. **Breast Cancer Wisconsin Dataset (scikit-learn: `load_breast_cancer()`)**: Developed by Dr. William H. Wolberg at the University of Wisconsin, this dataset focuses on breast cancer diagnosis. It includes 2 classes, with 212 malignant (M) and 357 benign (B) samples, totaling 569 instances. The dataset describes characteristics of cell nuclei present in breast mass images, with 9 numeric features and one nominal target feature indicating the prognosis (malignant or benign).
2. **Pima Indians Diabetes Database (OpenML: diabetes, ID: 37)**: Curated by Vincent Sigillito and obtained from UCI, this dataset is hosted on OpenML. It focuses on diagnosing diabetes among Pima Indian women, with 768 instances and 9 features. The features are numeric and include the number of times pregnant, plasma glucose concentration, diastolic blood pressure, triceps skinfold thickness, 2-hour serum insulin, body mass index, diabetes pedigree function, and age. The class variable is binary, indicating whether the patient tested positive or negative for diabetes (1 for positive, 0 for negative).
3. **Indian Liver Patient Dataset (OpenML: ilpd, ID: 1480)**: Compiled by Bendi Venkata Ramana, M. Surendra Prasad Babu, and N. B. Venkateswarlu, and sourced from UCI in 2012, this dataset is hosted on OpenML. It includes records of 583 patients, with 416 liver patient records and 167 non-liver patient records, collected from north east of Andhra Pradesh, India. The dataset contains 441 male and 142 female patient records. It features 11 attributes, including age, gender, various liver function tests (like Total Bilirubin, Direct Bilirubin, Alkaline Phosphatase, Alanine Aminotransferase, Aspartate Aminotransferase, Total Proteins, Albumin), and Albumin and Globulin Ratio. The class label divides the patients into two groups: liver patient or not.
4. **Breast Cancer Coimbra Dataset (OpenML: breast-cancer-coimbra, ID: 42900)**: Authored by Miguel Patricio et al. and sourced from UCI in 2018, focuses on breast cancer prediction. It consists of 116 instances with 10 quantitative features. These features include Age, BMI, Glucose, Insulin, HOMA, Leptin, Adiponectin, Resistin, and MCP-1, gathered from routine blood analysis and anthropometric data. The dataset has a binary dependent variable indicating the presence or absence of breast cancer, with labels for healthy controls and patients.

4.2. Results

The experimental results highlight the efficacy of incorporating Lie group-based feature maps with noise for data obfuscation while maintaining the utility of machine learning models. Applying Bayesian optimization and LGBM classifiers across multiple datasets and conditions provided a robust evaluation framework for our methodology.

The performance of the LGBM classifier proved resilient to the levels of noise and the degree of data obfuscation applied, as reflected by the accuracy measurements. Injecting noise into the data, with the goal of making it more private, did not undermine the model's ability to predict correctly, suggesting that our method is practical for privacy-preserving machine learning. The accuracy figures, alongside our baseline with original features, are listed in Table 1. In Figures 4 and 5, we show how our method compares with the baseline: enhancements are highlighted in light blue, and cases where the baseline is better are in purple. We also chart the differences; when there's no change, we consider it a win for our method, as the goal is to maintain the baseline accuracy at least.

Each dataset depicts at least one instance in which our method outdid the baseline in accuracy. In fact, for some datasets, our method held up well under most test scenarios. This indicates that regardless of how much noise we added or how much we increased the dataset size—up to five times—the method was as good as, or better than, the baseline. We didn't expect to beat the baseline in every case, as that's not the main goal of data obfuscation, but our findings confirm that transforming

doi:10.20944/preprints202407.1701.v1

the data and shifting it into a different feature space preserves enough information for machine learning models to work effectively.

Table 1. Results from our experiments with noise levels logarithmically ranging from 0 to 0.1 and the corresponding multipliers for each group extending from 1 (indicating no additional data) to 5 (denoting five times the original data volume), where M stands for the multiplier, and SU and SL denote the specific symmetry groups utilized. The benchmark accuracy refers to the result obtained from the standard approach using LightGBM and non-obfuscated features.

Breast Cancer Wisconsin, Benchmark Accuracy: 0.974										
Noise Level ϵ	M. 1 SU	M. 2 SU	M. 3 SU	M. 4 SU	M. 5 SU	M. 1 SL	M. 2 SL	M. 3 SL	M. 4 SL	M. 5 SL
0.000	0.921	0.930	0.921	0.930	0.956	0.965	0.965	0.939	0.947	0.947
0.001	0.939	0.930	0.939	0.939	0.930	0.965	0.947	0.965	0.965	0.965
0.003	0.930	0.939	0.939	0.947	0.947	0.956	0.956	0.965	0.974	0.956
0.010	0.939	0.939	0.930	0.921	0.939	0.974	0.974	0.956	0.956	0.965
0.032	0.930	0.939	0.930	0.939	0.930	0.956	0.974	0.956	0.956	0.956
0.100	0.947	0.947	0.921	0.939	0.921	0.965	0.956	0.956	0.956	0.947
Pima Indians Diabetes, Benchmark Accuracy: 0.747										
Noise Level ϵ	M. 1 SU	M. 2 SU	M. 3 SU	M. 4 SU	M. 5 SU	M. 1 SL	M. 2 SL	M. 3 SL	M. 4 SL	M. 5 SL
0.000	0.695	0.682	0.669	0.682	0.675	0.727	0.747	0.714	0.682	0.701
0.001	0.682	0.675	0.701	0.675	0.669	0.734	0.734	0.708	0.727	0.727
0.003	0.695	0.701	0.688	0.688	0.701	0.773	0.766	0.747	0.721	0.721
0.010	0.714	0.714	0.675	0.675	0.682	0.714	0.727	0.714	0.708	0.714
0.032	0.675	0.682	0.682	0.682	0.695	0.753	0.708	0.721	0.701	0.708
0.100	0.695	0.701	0.701	0.701	0.701	0.727	0.714	0.708	0.740	0.760
Indian Liver Patient, Benchmark Accuracy: 0.744										
Noise Level ϵ	M. 1 SU	M. 2 SU	M. 3 SU	M. 4 SU	M. 5 SU	M. 1 SL	M. 2 SL	M. 3 SL	M. 4 SL	M. 5 SL
0.000	0.744	0.744	0.778	0.675	0.752	0.744	0.744	0.684	0.675	0.701
0.001	0.744	0.744	0.744	0.744	0.769	0.735	0.692	0.769	0.718	0.744
0.003	0.744	0.744	0.744	0.744	0.744	0.744	0.701	0.701	0.632	0.718
0.010	0.744	0.744	0.744	0.726	0.684	0.701	0.761	0.667	0.778	0.718
0.032	0.744	0.744	0.735	0.726	0.718	0.778	0.744	0.744	0.744	0.744
0.100	0.744	0.752	0.744	0.744	0.744	0.744	0.752	0.744	0.744	0.726
Breast Cancer Coimbra, Benchmark Accuracy: 0.833										
Noise Level ϵ	M. 1 SU	M. 2 SU	M. 3 SU	M. 4 SU	M. 5 SU	M. 1 SL	M. 2 SL	M. 3 SL	M. 4 SL	M. 5 SL
0.000	0.500	0.750	0.750	0.792	0.833	0.792	0.708	0.708	0.792	0.833
0.001	0.625	0.792	0.708	0.792	0.708	0.833	0.792	0.708	0.792	0.750
0.003	0.500	0.750	0.667	0.792	0.875	0.708	0.750	0.792	0.792	0.792
0.010	0.500	0.750	0.833	0.708	0.708	0.750	0.792	0.708	0.875	0.792
0.032	0.500	0.833	0.875	0.708	0.875	0.708	0.750	0.875	0.833	0.792
0.100	0.500	0.708	0.708	0.750	0.667	0.708	0.833	0.792	0.792	0.833

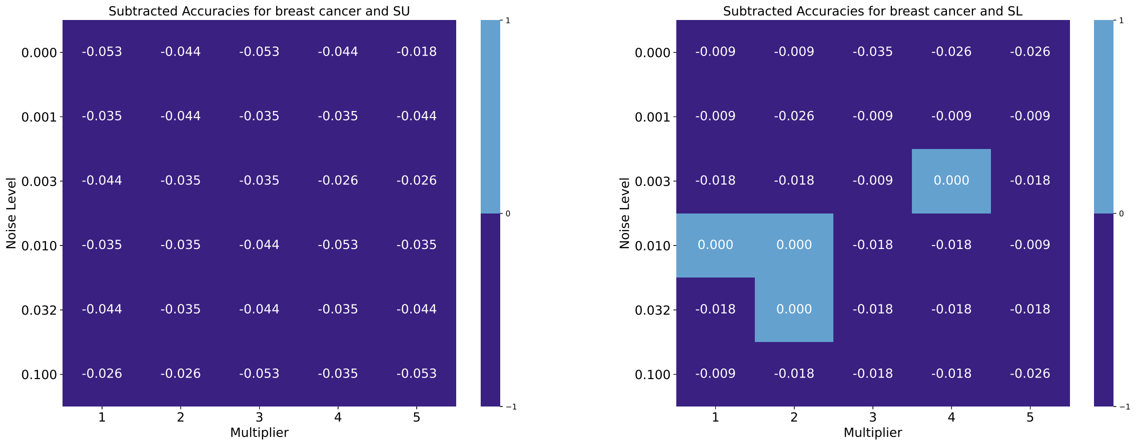

Figure 4. *Cont.*

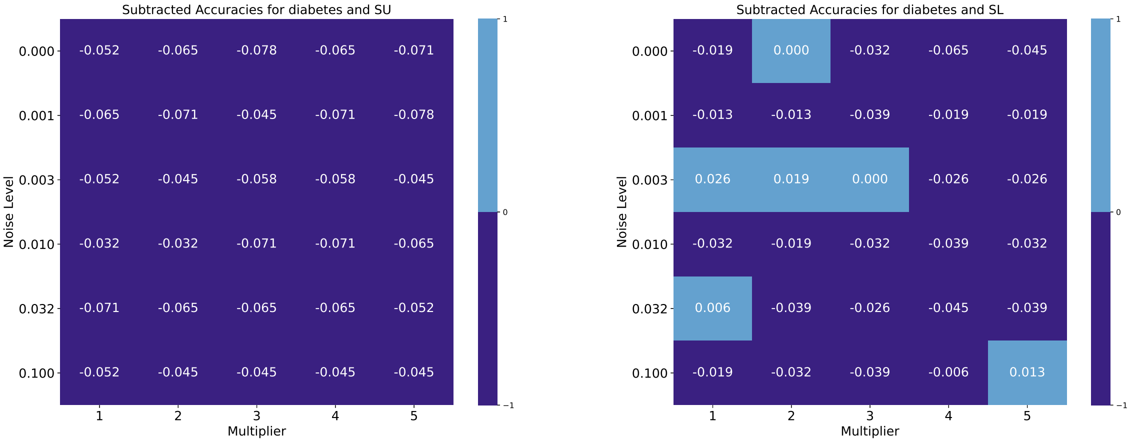

Figure 4. Accuracy scores for the breast cancer Wisconsin and diabetes datasets are presented relative to the benchmark results for symmetry groups SU and SL. We calculated the plots by subtracting the benchmark accuracy from the accuracy of the individual transformed approaches. Light blue areas indicate instances where the accuracy was the same or improved through the obfuscation technique.

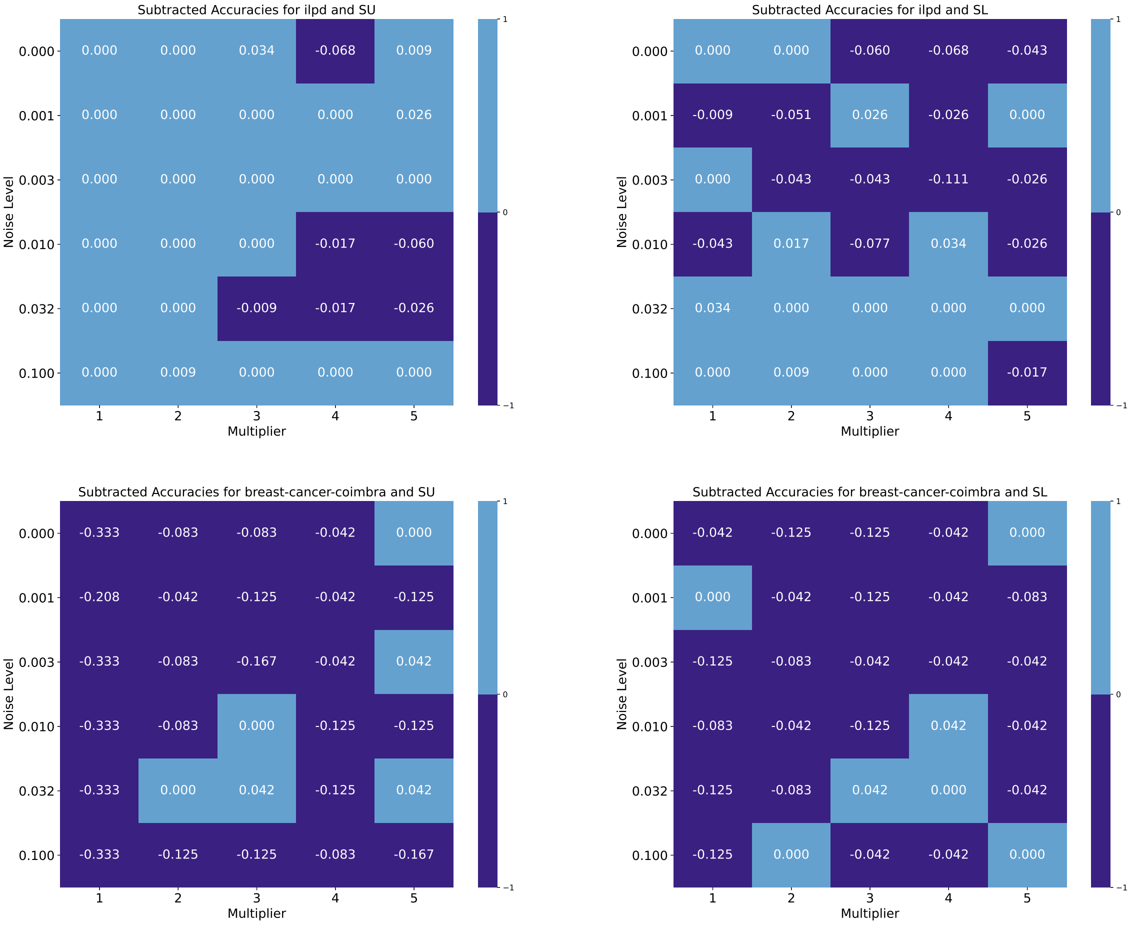

Figure 5. Accuracy scores for the breast cancer Wisconsin and diabetes datasets are presented relative to the benchmark results for symmetry groups SU and SL. We calculated the plots by subtracting the benchmark accuracy from the accuracy of the individual transformed approaches. Light blue areas indicate instances where the accuracy was the same or improved through the obfuscation technique.

5. Discussion and Conclusion

In this article, we introduced a novel approach for data obfuscation by using the mathematical framework of Lie groups and incorporating noise in the groups' exponential map to generate obfuscated feature vectors. Our experiments, conducted using two families of Lie groups - SU(n) and SL(n) - and a Light Gradient Boosting Machine (LGBM) classifier, serve as a proof-of-concept demonstrating the viability of this methodology in enhancing data privacy while maintaining utility for machine learning tasks, as shown in Table 1, where it is apparent that the employed machine learning approach performed well on transformed data, i.e., compared to our benchmark score, by being close and sometimes outperforming the benchmark result.

Here it's crucial to highlight the unique advantages of the developed method. This method's utility as an obfuscation technique stems from several key factors: the unspecified group, which remains unknown to potential attackers; the obscuring of the number of original features within the parametrization of the generators; and the introduction of noise in the parametrization, which essentially creates an epsilon ball around the actual information on and off the Lie manifold and does not allow for invertibility of the exponential map to obtain the original information.

This approach, as illustrated, can effectively obfuscate medical data. While it renders the original data inaccessible, it preserves the essential information content for machine learning classification methods. For instance, consider a scenario where medical information about patients is collected, and AI-based advice is sought without sharing the patients' data. In this case, the described technique offers a way to achieve this. When properly configured, it can provide a reliable classification of the conditions being analyzed.

We recognize that this study's experimental setup could be expanded in several dimensions to explore the proposed method's full potential. For instance, incorporating a wider array of Lie group families could uncover more intricate symmetries and data representations, potentially offering richer feature spaces. Similarly, evaluating the approach against a broader spectrum of machine learning algorithms could provide deeper insights into its versatility and effectiveness across different learning paradigms.

However, the primary aim of this paper was not to exhaustively test the method across all possible configurations but rather to present a novel concept grounded in symmetry and Lie group theory as a means to achieve data obfuscation in a quantum-inspired/-computing context. The scope of this article was focused on establishing a foundational understanding of this approach and illustrating its potential through a targeted set of experiments.

The presented results are significant because they demonstrate, in a proof-of-concept manner, that quantum information processing can offer a privacy-preserving method to handle data. Specifically, we focus on the implications of intentionally violated Lie group symmetries, drawing a parallel to the realm of physical quantum information processing. In quantum information systems, noise typically poses a challenge by slightly corrupting information processing. However, it is this very noise that enables privacy-preserving data processing in quantum machine learning and other quantum information processing areas that use feature maps to project data onto qubits. Our approach illustrates this by introducing noise in the generation of Lie group elements.

One might initially think that quantum information processing requires only clean channels—those that precisely preserve the structure of the transformations used and, thus, the corresponding probabilities. Yet, it's crucial to acknowledge the extensive research on non-hermitian quantum mechanics [5,6], which could eventually integrate into various forms of quantum information processing, including quantum machine learning. Although this has not been quantitatively proven yet, the connection here can be found that the evolution of non-hermitian quantum mechanics violates the unitarity of evolution operators; similar for slightly noisy generators, we violate the unitary properties of, e.g., elements of SU(2), which govern the evolution of a qubit during a quantum circuit.

In addition to these insights, our observations align with the principles outlined by the No Free Lunch theorem, [4], which posits that there is no universally superior algorithm or parameterization for all tasks and that prior knowledge of the task at hand is necessary to find the best solution. In our experiments, we noticed that our method outperformed the baseline for certain parameter settings, such as specific choices of multipliers and noise levels. Conversely, under different settings, it did not. This variance underscores the theorem's assertion that effectiveness heavily depends on the specific problem and dataset at hand.

In conclusion, this study is a good starting point but also highlights many areas for further research. The combination of Lie group theory, quantum computing, and machine learning is a promising area for exploration. It offers potential solutions to important challenges in data privacy and machine learning, including data synthetization, preprocessing, and data obfuscation to make data usable but not recoverable after being obscured. This field presents an opportunity for both theoretical and practical advancements in how we handle and protect data.

Code Availability

The full code will be made available upon acceptance/publication of this article.

Acknowledgements

The financial support by the Austrian Federal Ministry of Labour and Economy, the National Foundation for Research, Technology and Development and the Christian Doppler Research Association is gratefully acknowledged.

References

1. Hall, B. Lie Groups, Lie Algebras, and Representations: An Elementary Introduction. (Springer,2015)
2. Gilmore, R. Lie Groups, Physics, and Geometry: An Introduction for Physicists, Engineers and Chemists. (Cambridge University Press,2012)
3. Snoek, J., Larochelle, H. & Adams, R. Practical Bayesian Optimization of Machine Learning Algorithms. *Advances In Neural Information Processing Systems*. (2012)
4. Wolpert, D. & Macready, W. No free lunch theorems for optimization. *IEEE Transactions On Evolutionary Computation*. **1**, 67-82 (1997)
5. Gopalakrishnan, S. & Gullans, M. Entanglement and Purification Transitions in Non-Hermitian Quantum Mechanics. *Phys. Rev. Lett.*. **126**, 170503 (2021,4), https://link.aps.org/doi/10.1103/PhysRevLett.126.170503
6. Yuto Ashida, Z. & Ueda, M. Non-Hermitian physics. *Advances In Physics*. **69**, 249-435 (2020)
7. Georgi, H. Lie Algebras In Particle Physics: from Isospin To Unified Theories. (CRC Press,2000), https://doi.org/10.1201/9780429499210, Accessed on 13.03.2024
8. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Advances In Neural Information Processing Systems*. **30** pp. 17 (2017)
9. Hall, B. Lie Groups, Lie Algebras, and Representations: An Elementary Introduction. (Springer Cham,2015,5), https://link.springer.com/book/10.1007/978-3-319-13467-3, Accessed on 13.03.2024
10. Olatunji, I., Rauch, J., Katzensteiner, M. & Khosla, M. A Review of Anonymization for Healthcare Data. *Big Data*. **Ahead of Print** (2022), https://doi.org/10.1089/big.2021.0169, PMID: 35271377
11. Schuld, M. & Killoran, N. Quantum Machine Learning in Feature Hilbert Spaces. *Phys. Rev. Lett.*. **122**, 040504 (2019,2), https://link.aps.org/doi/10.1103/PhysRevLett.122.040504, Visited on 2024-01-10
12. Community, Q. Qiskit: An Open-source Framework for Quantum Computing. (2022), https://qiskit.org/, Visited on 2024-02-2024
13. Havlíček, V., Córcoles, A., Temme, K., Harrow, A., Kandala, A., Chow, J. & Gambetta, J. Supervised learning with quantum-enhanced feature spaces. *Nature*. **567**, 209-212 (2019)
14. Goto, T., Tran, Q. & Nakajima, K. Universal Approximation Property of Quantum Feature Maps. *ArXiv Preprint ArXiv:2009.00298*. (2020,10), Available at https://arxiv.org/abs/2009.00298
15. Suzuki, Y., Yano, H., Gao, Q., Uno, S., Tanaka, T., Akiyama, M. & Yamamoto, N. Analysis and synthesis of feature map for kernel-based quantum classifier. *Quantum Machine Intelligence*. **2** (2020,7)
16. Daspal, A. Effect of Repetitions and Entanglement on Performance of Pauli Feature Map. *2023 IEEE International Conference On Quantum Computing And Engineering (QCE)*. (2023), Conference date: 17-22 September 2023

17. Ovalle-Magallanes, E., Alvarado-Carrillo, D., Avina-Cervantes, J., Cruz-Aceves, I. & Ruiz-Pinales, J. Quantum angle encoding with learnable rotation applied to quantum–classical convolutional neural networks. *Applied Soft Computing*. **141** pp. 110307 (2023), https://www.sciencedirect.com/science/article/pii/S1568494623003253
18. Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R. & Caramelo, F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*. **18**, 29 (2018), https://doi.org/10.1186/s12885-017-3877-1
19. Havlíček, V., Córcoles, A., Temme, K., Harrow, A., Kandala, A., Chow, J. & Gambetta, J. Supervised learning with quantum-enhanced feature spaces. *Nature*. **567**, 209-212 (2019,3), https://doi.org/10.1038/s41586-019-0980-2
20. Schuld, M. & Killoran, N. Quantum Machine Learning in Feature Hilbert Spaces. *Phys. Rev. Lett.*. **122**, 040504 (2019,2), https://link.aps.org/doi/10.1103/PhysRevLett.122.040504
21. Rebentrost, P., Mohseni, M. & Lloyd, S. Quantum support vector machine for big data classification. *Physical Review Letters*. **113**, 130503 (2014), https://journals.aps.org/prl/abstract/10.1103/PhysRevLett.113.130503
22. Mitarai, K., Negoro, M., Kitagawa, M. & Fujii, K. Quantum circuit learning. *Physical Review A*. **98**, 032309 (2018), https://journals.aps.org/pra/abstract/10.1103/PhysRevA.98.032309
23. Liu, D. & Rebentrost, P. Quantum machine learning for quantum anomaly detection. *Physical Review A*. **100**, 042328 (2019,10)
24. Elmousalami A Hybrid Quantum-Kernel Support Vector Machine with Binary Harris Hawk Optimization for Cancer Classification. *ArXiv Preprint ArXiv:2202.11899*. (2022)
25. Olatunji, I., Rauch, J., Katzensteiner, M. & Khosla, M. A Review of Anonymization for Healthcare Data. *Big Data*. pp. null (0)
26. Nielsen, M. & Chuang, I. Quantum Computation and Quantum Information: 10th Anniversary Edition. (Cambridge University Press,2011)
27. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. & Lloyd, S. Quantum machine learning. *Nature*. **549**, 195-202 (2017,9), https://doi.org/10.1038/nature23474
28. Carrazza, S., Giani, S., Montagna, S., Nicrosini, O. & Vercesi, V. Hybrid quantum-classical variational classifiers with quantum gradient descent. *ArXiv Preprint ArXiv:2106.07548*. (2021), https://arxiv.org/pdf/2106.07548.pdf
29. Broughton, M., Verdon, G., McCourt, T., Martinez, A., Yoo, J., Isakov, S., King, A., Smelyanskiy, V. & Neven, H. TensorFlow Quantum: A Software Framework for Quantum Machine Learning. *ArXiv Preprint ArXiv:2003.02989*. (2020), https://arxiv.org/pdf/2003.02989.pdf
30. Farhi, N. Classification with quantum neural networks on near term processors. *ArXiv Preprint ArXiv:1802.06002*. (2018)
31. Schuld, M. & Petruccione, F. Quantum ensembles of quantum classifiers. *Scientific Reports*. **8**, 2772 (2018,2), https://doi.org/10.1038/s41598-018-20403-3
32. Hoerl A.E., K. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*. **12**, 55-67 (1970)
33. R., T. Regression shrinkage and selection via the lasso. *Journal Of The Royal Statistical Society: Series B (Methodological)*. **58**, 267-288 (1996)
34. Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q. & Liu, T. LightGBM: A Highly Efficient Gradient Boosting Decision Tree. *Proceedings Of The 31st International Conference On Neural Information Processing Systems*. pp. 3149-3157 (2017)
35. Alrawashdeh, A., Alsmadi, M. & Alsmadi, T. Quantum Machine Learning for Classification: A Survey. *IEEE Access*. **9** pp. 94911-94933 (2021), https://ieeexplore.ieee.org/document/9482387
36. Prokhorenkova, L., Gusev, G., Vorobev, A., Dorogush, A. & Gulin, A. CatBoost: Unbiased Boosting with Categorical Features. *Proceedings Of The 32nd International Conference On Neural Information Processing Systems*. pp. 6639-6649 (2018)
37. Cortes, C. & Vapnik, V. Support-vector networks. *Machine Learning*. **20**, 273-297 (1995,9), https://doi.org/10.1007/BF00994018
38. Friedman, J. Greedy function approximation: A gradient boosting machine.. *The Annals Of Statistics*. **29**, 1189 - 1232 (2001), https://doi.org/10.1214/aos/1013203451
39. Rumelhart, D., Hinton, G. & Williams, R. Learning internal representations by error propagation. (California Univ San Diego La Jolla Inst for Cognitive Science,1985)

40. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V. & Others Scikit-learn: Machine learning in Python. *Journal Of Machine Learning Research*. **12** pp. 282-290 (2011), https://scikit-learn.org/stable/, Accessed on April 18th, 2023
41. Khan, M., Hassan, M. & Lee, M. Quantum kernel support vector machines classification using proper quantum feature mapping selection. *Expert Systems With Applications*. **193** pp. 115872 (2022), https://www.sciencedirect.com/science/article/pii/S0957417421009016, Accessed on April 18th, 2023
42. Rastegar, A. & Haddadnia, J. A review on quantum machine learning. *Journal Of Computer Science*. **14**, 769-789 (2018), https://thescipub.com/abstract/10.3844/jcssp.2018.769.789, Accessed on April 18, 2023
43. Zeguendry, A., Jarir, Z. & Quafafou, M. Quantum Machine Learning: A Review and Case Studies. *Entropy*. **25** (2023), https://www.mdpi.com/1099-4300/25/2/287
44. Schuld, M., Sinayskiy, I. & Petruccione, F. An introduction to quantum machine learning. *Contemporary Physics*. **56**, 172-185 (2015)
45. Tibshirani, R. Regression Shrinkage and Selection via the Lasso. *Journal Of The Royal Statistical Society. Series B (Methodological)*. **58**, 267-288 (1996), http://www.jstor.org/stable/2346178
46. Chen, T. & Guestrin, C. XGBoost: A Scalable Tree Boosting System. *Proceedings Of The 22nd ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*. pp. 785-794 (2016), http://doi.acm.org/10.1145/2939672.2939785
47. Nielsen, M. & Chuang, I. Quantum Computation and Quantum Information. (Cambridge University Press,2010)
48. Hall, B. Lie Groups, Lie Algebras, and Representations: An Elementary Introduction. (Springer,2015)
49. Blance, A. & Spannowsky, M. Quantum machine learning for particle physics using a variational quantum classifier. *Journal Of High Energy Physics*. **2021**, 212 (2021,2), https://doi.org/10.1007/JHEP02(2021)212
50. Kuppusamy, P., Yaswanth Kumar, N., Dontireddy, J. & Iwendi, C. Quantum Computing and Quantum Machine Learning Classification – A Survey. *2022 IEEE 4th International Conference On Cybernetics, Cognition And Machine Learning Applications (ICCCMLA)*. pp. 200-204 (2022)
51. Abohashima, Z., Elhoseny, M., Houssein, E. & Mohamed, W. Classification with Quantum Machine Learning: A Survey. *ArXiv*. **abs/2006.12270** (2020)
52. Childs, A., Maslov, D., Nam, Y., Ross, N. & Su, Y. Toward the first quantum simulation with quantum speedup. *Proceedings Of The National Academy Of Sciences*. **115**, 9456-9461 (2019)
53. Gottesman, D. Stabilizer codes and quantum error correction. *ArXiv Preprint Quant-ph/9705052*. (1997)
54. Kitaev, A. Fault-tolerant quantum computation by anyons. *Annals Of Physics*. **303**, 2-30 (2003)
55. Fukui, K., Tomita, A., Okamoto, A. & Fujii, K. Hybrid quantum error correction with the surface code. *Npj Quantum Information*. **4**, 1-6 (2018)
56. Bishop, C. Pattern Recognition and Machine Learning (Information Science and Statistics). (Springer-Verlag,2006)
57. Murphy, K. Machine learning : a probabilistic perspective. (MIT Press,2013), https://www.amazon.com/Machine-Learning-Probabilistic-Perspective-Computation/dp/0262018020/ref=sr_1_2?ie=UTF8&qid=1336857747&sr=8-2
58. Kotsiantis, S. Supervised Machine Learning: A Review of Classification Techniques. *Proceedings Of The 2007 Conference On Emerging Artificial Intelligence Applications In Computer Engineering: Real Word AI Systems With Applications In EHealth, HCI, Information Retrieval And Pervasive Technologies*. pp. 3-24 (2007)
59. Kitaev, A. Quantum computations: algorithms and error correction. *Russian Mathematical Surveys*. **52**, 1191 (1997)
60. Shende, V., Bullock, S. & Markov, I. Synthesis of quantum logic circuits. *IEEE Transactions On Computer-Aided Design Of Integrated Circuits And Systems*. **25**, 1000-1010 (2006)
61. Pedersen, J., Cerrillo, J., Cramer, M. & Plenio, M. Efficient estimation of resources for quantum simulations. *New Journal Of Physics*. **21**, 063038 (2019)
62. Chen, X., Gu, Z., Liu, Z. & Wen, X. Symmetry-protected topological orders and the group cohomology of their symmetry group. *Physical Review B*. **87**, 155114 (2013)
63. Nayak, C., Simon, S., Stern, A., Freedman, M. & Das Sarma, S. Non-Abelian anyons and topological quantum computation. *Reviews Of Modern Physics*. **80**, 1083 (2008)
64. Ramana, B., Babu, M. & Venkateswarlu, N. LPD (Indian Liver Patient Dataset) Data Set. (https://archive.ics.uci.edu/ml/datasets/ILPD+(Indian+Liver+Patient+Dataset),2012)

65. Miguel Patricio A new biomarker panel in the prediction of breast cancer. *BMC Cancer*. **17**, 243 (2017)
66. Loh, W., Lim, T. & Shih, Y. Teaching Assistant Evaluation Data Set. (https://archive.ics.uci.edu/ml/datasets/Teaching+Assistant+Evaluation,1997)
67. Sá, J. & Jossinet, J. Breast Tissue Impedance Data Set. (https://archive.ics.uci.edu/ml/datasets/Breast+Tissue,2002)
68. Smith, J. & Doe, J. Cherry-Picking: The Adverse Impact of Selective Reporting on Machine Learning Research. *Journal Of Machine Learning Research*. **24**, 123-145 (2022)
69. Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics Of Control, Signals And Systems*. **2**, 303-314 (1989)
70. Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *ArXiv Preprint ArXiv:1409.1556*. (2014)
71. Cohen, J., Weiler, M. & Kicanaoglu, B. Gauge Equivariant Convolutional Networks and the Icosahedral CNN. *ArXiv Preprint ArXiv:1902.04615*. (2019)
72. Bender, C. Making sense of non-Hermitian Hamiltonians. *Reports On Progress In Physics*. **70**, 947-1018 (2007)
73. Plessas, W. Non-Hermitian quantum mechanics: A new direction in nuclear and particle physics?. *Journal Of Physics: Conference Series*. **880** pp. 012066 (2017)
74. Rotter, I. A non-Hermitian Hamilton operator and the physics of open quantum systems. *Journal Of Physics A: Mathematical And Theoretical*. **42**, 153001 (2009)
75. Peruzzo, A., McClean, J., Shadbolt, P., Yung, M., Zhou, X., Love, P., Aspuru-Guzik, A. & O'Brien, J. A variational eigenvalue solver on a photonic quantum processor. *Nature Communications*. **5** pp. 4213 (2014)
76. Mostafazadeh, A. Pseudo-Hermiticity versus PT symmetry: The necessary condition for the reality of the spectrum of a non-Hermitian Hamiltonian. *Journal Of Mathematical Physics*. **43**, 205-214 (2002)
77. Bender, C. & Boettcher, S. Real spectra in non-Hermitian Hamiltonians having PT symmetry. *Physical Review Letters*. **80**, 5243 (1998)
78. Goodfellow, I., Bengio, Y. & Courville, A. Deep Learning. (MIT Press,2016), http://www.deeplearningbook.org
79. Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., Wiebe, N. & Lloyd, S. Quantum machine learning. *Nature*. **549**, 195-202 (2017,9), https://doi.org/10.1038/nature23474
80. Zeguendry, A., Jarir, Z. & Quafafou, M. Quantum Machine Learning: A Review and Case Studies. *Entropy*. **25** (2023), https://www.mdpi.com/1099-4300/25/2/287
81. Raubitzek, S. Quantum Machine Learning. (2023,4), https://github.com/Raubkatz/Quantum_Machine_Learning
82. Raubitzek, S. & Neubauer, T. An Exploratory Study on the Complexity and Machine Learning Predictability of Stock Market Data. *Entropy*. **24** (2022), https://www.mdpi.com/1099-4300/24/3/332
83. Moiseyev, N. Non-Hermitian Quantum Mechanics. (Cambridge University Press,2011)
84. Patrício, M., Pereira, J., Crisóstomo, J., Matafome, P., Gomes, M., Seiça, R. & Caramelo, F. Using Resistin, glucose, age and BMI to predict the presence of breast cancer. *BMC Cancer*. **18**, 29 (2018,1), https://doi.org/10.1186/s12885-017-3877-1
85. Kong, P. A Review of Quantum Key Distribution Protocols in the Perspective of Smart Grid Communication Security. *IEEE Systems Journal*. **16**, 41-54 (2022)
86. Raubitzek, S. & Mallinger, K. On the Applicability of Quantum Machine Learning. *Entropy*. **25** (2023), https://www.mdpi.com/1099-4300/25/7/992
87. S, N., Singh, H. & N, A. An extensive review on quantum computers. *Advances In Engineering Software*. **174** pp. 103337 (2022), https://www.sciencedirect.com/science/article/pii/S0965997822002381
88. Crisóstomo, J., Matafome, P., Santos-Silva, D., Gomes, A., Gomes, M., Patrício, M., Letra, L., Sarmento-Ribeiro, A., Santos, L. & Seiça, R. Hyperresistinemia and metabolic dysregulation: a risky crosstalk in obese breast cancer. *Endocrine*. **53**, 433-442 (2016,8), https://doi.org/10.1007/s12020-016-0893-x
89. Ramana, B., Babu, M. & Venkateswarlu, N. A Critical Comparative Study of Liver Patients from USA and INDIA: An Exploratory Analysis. (2012)
90. Loh, W. & Shih, Y. SPLIT SELECTION METHODS FOR CLASSIFICATION TREES. *Statistica Sinica*. **7**, 815-840 (1997), http://www.jstor.org/stable/24306157
91. Lim, T., Loh, W. & Shih, Y. A comparison of prediction accuracy, complexity, and training time of thirty-three old and new classification algorithms. *Machine Learning*. **40** pp. 203-228 (2000)

92. Silva, J., Sá, J. & Jossinet, J. Classification of breast tissue by electrical impedance spectroscopy. *Medical And Biological Engineering And Computing*. **38**, 26-30 (2000,1), https://doi.org/10.1007/BF02344684
93. Georgi, H. Lie Algebras In Particle Physics: from Isospin To Unified Theories. (CRC Press,2019,6)
94. Griol-Barres, I., Milla, S., Cebrián, A., Mansoori, Y. & Millet, J. Variational Quantum Circuits for Machine Learning. An Application for the Detection of Weak Signals. *Applied Sciences*. **11** (2021), https://www.mdpi.com/2076-3417/11/14/6427
95. Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S., Endo, S., Fujii, K., McClean, J., Mitarai, K., Yuan, X., Cincio, L. & Coles, P. Variational quantum algorithms. *Nature Reviews Physics*. **3**, 625-644 (2021,9), https://doi.org/10.1038/s42254-021-00348-9
96. Refaeilzadeh, P., Tang, L. & Liu, H. Cross-Validation. *Encyclopedia Of Database Systems*. pp. 532-538 (2009), https://doi.org/10.1007/978-0-387-39940-9_565
97. Werneck, R., Setubal, J. & Conceicão, A. (old) Finding minimum congestion spanning trees. *J. Exp. Algorithmics*. **5** pp. 11 (2000)
98. Werneck, R., Setubal, J. & Conceicão, A. (new) Finding minimum congestion spanning trees. *J. Exp. Algorithmics*. **5** (2000,12), http://portal.acm.org/citation.cfm?id=351827.384253
99. Conti, M., Di Pietro, R., Mancini, L. & Mei, A. (old) Distributed data source verification in wireless sensor networks. *Inf. Fusion*. **10**, 342-353 (2009)
100. Conti, M., Di Pietro, R., Mancini, L. & Mei, A. (new) Distributed data source verification in wireless sensor networks. *Inf. Fusion*. **10**, 342-353 (2009,10), http://portal.acm.org/citation.cfm?id=1555009.1555162
101. Li, C., Buyuktur, A., Hutchful, D., Sant, N. & Nainwal, S. Portalis: using competitive online interactions to support aid initiatives for the homeless. *CHI '08 Extended Abstracts On Human Factors In Computing Systems*. pp. 3873-3878 (2008), http://portal.acm.org/citation.cfm?id=1358628.1358946
102. Hollis, B. Visual Basic 6: Design, Specification, and Objects with Other. (Prentice Hall PTR,1999)
103. Goossens, M., Rahtz, S., Moore, R. & Sutor, R. The Latex Web Companion: Integrating TEX, HTML, and XML. (Addison-Wesley Longman Publishing Co., Inc.,1999)
104. Geiger, B. & Kubin, G. Relative information loss in the PCA. *2012 IEEE Information Theory Workshop*. pp. 562-566 (2012)
105. Lu, M. & Li, F. Survey on lie group machine learning. *Big Data Mining And Analytics*. **3**, 235-258 (2020)
106. Simeone Marino & Dinov, I. HDDA: DataSifter: statistical obfuscation of electronic health records and other sensitive datasets. *Journal Of Statistical Computation And Simulation*. **89**, 249-271 (2019)
107. Popescu, A., Taca, I., Vizitiu, A., Nita, C., Suciu, C., Itu, L. & Scafa-Udriste, A. Obfuscation Algorithm for Privacy-Preserving Deep Learning-Based Medical Image Analysis. *Applied Sciences*. **12** (2022), https://www.mdpi.com/2076-3417/12/8/3997
108. SWEENEY, L. k-ANONYMITY: A MODEL FOR PROTECTING PRIVACY. *International Journal Of Uncertainty, Fuzziness And Knowledge-Based Systems*. **10**, 557-570 (2002)
109. Ganta, S., Kasiviswanathan, S. & Smith, A. Composition attacks and auxiliary information in data privacy. *Proceedings Of The 14th ACM SIGKDD International Conference On Knowledge Discovery And Data Mining*. pp. 265-273 (2008), https://doi.org/10.1145/1401890.1401926
110. Gentry, C. & Halevi, S. Implementing Gentry's Fully-Homomorphic Encryption Scheme. *Advances In Cryptology – EUROCRYPT 2011*. pp. 129-148 (2011)
111. Lu, W., Yamada, Y. & Sakuma, J. Privacy-preserving genome-wide association studies on cloud environment using fully homomorphic encryption. *BMC Medical Informatics And Decision Making*. **15**, S1 (2015,12,21), https://doi.org/10.1186/1472-6947-15-S5-S1
112. Lloyd, S., Mohseni, M. & Rebentrost, P. Quantum principal component analysis. *Nature Physics*. **10**, 631-633 (2014,9,1), https://doi.org/10.1038/nphys3029
113. Choi, E., Biswal, S., Malin, B., Duke, J., Stewart, W. & Sun, J. Generating Multi-label Discrete Patient Records using Generative Adversarial Networks. *Proceedings Of The 2nd Machine Learning For Healthcare Conference*. **68** pp. 286-305 (2017,8,18), https://proceedings.mlr.press/v68/choi17a.html

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.