Hindawi
Computational Intelligence and Neuroscience
Volume 2022, Article ID 5314671, 9 pages
https://doi.org/10.1155/2022/5314671

Research Article

Leakage Prediction in Machine Learning Models When Using Data from Sports Wearable Sensors

Qizheng Dong

Zhengzhou University of Science and Technology, Zhengzhou, Henan 450000, China

Correspondence should be addressed to Qizheng Dong; dongqizheng1982@126.com

Received 1 April 2022; Revised 19 April 2022; Accepted 25 April 2022; Published 17 May 2022

Academic Editor: Konstantinos Demertzis

Copyright © 2022 Qizheng Dong. This is an open access article distributed under the Creative Commons Attribution License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

One of the major problems in machine learning is data leakage, which can be directly related to adversarial type attacks, raising serious concerns about the validity and reliability of artificial intelligence. Data leakage occurs when the independent variables used to teach the machine learning algorithm include either the dependent variable itself or a variable that contains clear information that the model is trying to predict. This data leakage results in unreliable and poor predictive results after the development and use of the model. It prevents the model from generalizing, which is required in a machine learning problem and thus causes false assumptions about its performance. To have a solid and generalized forecasting model, which will be able to produce remarkable forecasting results, we must pay great attention to detecting and preventing data leakage. This study presents an innovative system of leakage prediction in machine learning models, which is based on Bayesian inference to produce a thorough approach to calculating the reverse probability of unseen variables in order to make statistical conclusions about the relevant correlated variables and to calculate accordingly a lower limit on the marginal likelihood of the observed variables being derived from some coupling method. The main notion is that a higher marginal probability for a set of variables suggests a better fit of the data and thus a greater likelihood of a data leak in the model. The methodology is evaluated in a specialized dataset derived from sports wearable sensors.

1. Introduction

Machine learning models typically receive input data and solve problems such as pattern recognition by applying a sequence of particular transformations. The majority of these transformations turn out to be extremely sensitive to modest changes in input. Under specific scenarios, using this sensitivity can result in a difference in the behavior of the learning algorithm [1, 2]. Adversarial attack is the design of an adequate input in a specific way that leads the learning algorithm to erroneous outputs while not easily noticed by human observers. It is a severe concern in the reliability and security of artificial intelligence technologies. The issue arises because learning techniques are intended for use in stable situations where training and test data are generated from the same, possibly unknown distribution [3]. A trained neural network, for example, represents a significant decision limit corresponding to a standard class. Of course, the restriction is not without flaws. A correctly designed and implemented attack, which corresponds to a modified input form a slightly differentiated dataset, can cause the algorithm to make an incorrect judgment (wrong class) [4–6].

Developing and selecting machine learning methodologies to solve complex, usually nonlinear, problems is inextricably linked to the area of application and the target problem it seeks to solve. This is one of the essential processes of preprocessing the area of interest and the dataset, as the choice of appropriate algorithms depends on not only the nature and dynamics of the problem but also the characteristics of the available data, such as volume, number, and type of variables in question. The preprocessing of the data concerns the tests and the preparation work that should be carried out in the examined dataset before the use and application of machine learning algorithms. This method is critical because if the quality of usage or training data is not

ensured, the algorithms' performance will be subpar or the algorithms may produce false results [6, 7].

In general, data preparation/preprocessing entails dealing with scenarios when the original data have issues such as contradicting information, coding discrepancies, field terminology, and units of measurement. However, more critical issues such as the presence of lost values, noise, and extreme values and dealing with special requirements that necessitate data transformation, such as discretization, normalization, dimension reduction, or the selection of the most appropriate features, must be addressed [9–11]. It should be noted that several techniques can be used in preprocessing processes, with the choice of the best strategy arising from the nature of the field of knowledge, the problem to be addressed, the available data, and the machine learning algorithm used.

One of the most critical errors that occur during the preprocessing of data for use by machine learning algorithms is data leakage. The leak in question refers to cases where, inadvertently or even intentionally, the value that the model wishes to predict (dependent variable) is contained indirectly or directly in the features that are called to train the algorithm (independent variables). Any variable that provides transparent information about the value that the model is trying to predict is considered a data leak and leads to fictitious results. An obvious solution to this problem is to apply preprocessing only to the training set. Using preprocessing techniques to the whole dataset will make the model learn the training and the test sets, resulting in a data leak, and thus the model fails to generalize [2, 12, 13].

The major problem of data leakage occurs when there is a severe indirect interaction of features which is not easy to detect. It is, for example, a widespread phenomenon in machine learning experiments; the relationship between the dependent and the independent variable is complex (e.g., polynomial, trigonometric, and so on), so new features may be created that seem to help capture this relationship. Still, in practice, they create serious data leaks [14, 15].

Similarly, combinations may exist between independent and dependent variables through, for example, an arithmetic operation, a modification, or a conversion to make them more important in explaining the discrepancies in the data than if they remained separate. Creating a new opportunity through the interaction of existing features creates data leaks and significant bias in the final machine learning model [4, 7, 11].

For example, Lu et al. [15] developed a weighted context graph model (WCGM) for information leakage, with the critical goals of first increasing the contextual relevance of information, second classifying the tested data based on the commonality characteristics of its context graphs, and third preserving data proprietors' privacy. The weighted context network reduces complexity by using key sensitive phrases as nodes and contextual linkages as edges. The proposed maximum subgraph matching approach and deep learning algorithms are used to evaluate the similarity of the tested information and the pattern, as well as the responsiveness of the tested data to match the converted data better. The proposed model surpassed the competition regarding accuracy, recall, and run time, indicating its ability to detect real-time data leaks.

Using a variety of datasets, Salem et al. [14] provided research on the new and developing danger of membership inference attacks, demonstrating the efficacy of the suggested assaults across sectors. They offer two defensive strategies to alleviate the problem. The first, known as dropout, involves randomly deleting specific nodes in each fully linked neural system training step. In contrast, the second, known as model stacking, involves organizing numerous ML models in a ranked order [16]. Extensive testing has shown that our defensive strategies may significantly lower the performance of a membership inference attempt while retaining a high degree of usefulness, i.e., good target model prediction accuracy. They also suggest a defensive mechanism against a larger class of inclusion inference assaults while maintaining the ML model's high usefulness.

In this work, we proposed an innovative system of leakage prediction in machine learning models, which calculates a lower limit for the marginal probability of the observed variables coming from a coupling method, which shows that in an examined machine learning model, there is data leakage. The methodology is implemented based on the Bayesian inference methodology [17–19]. The model's goal is to generate an analytical approach to the reverse probability of unobserved variables [20, 21], to draw statistical inferences about the important correlated variables, and to compute a lower limit for the marginal likelihood of observable variables generated from a coupling method. The highest probability indicates that there is a data leak [22]. This is done to have a solid and generalized forecasting model, which will produce remarkable forecasting results without data leakages.

2. Proposed Approach

The proposed implementation is based on Bayesian inference [23–25], which is a method of approaching intractable problems that arise in highly fuzzy environments. More specifically, the methodology offers a secure solution for the observed variables and unknown parameters and latent states of variables, characterized by different types of relationships (interconnected, transformed, hidden, random, and so on). A prior distribution, a posterior distribution, and a likelihood function are used to illustrate Bayesian inference [26] in Figure 1.

The prediction error is defined as the difference between the previous expectation and the likelihood function's peak (i.e., reality). The variance of the prior is the source of uncertainty. The variance of the likelihood function is referred to as noise [27].

Parameters and latent variables are grouped as "unobserved variables." So, with the proposed method, the purpose is as follows [28–31]:

(1) In order to generate an analytical approach to the reverse probability of unobserved variables, develop statistical findings for the important correlated variables.

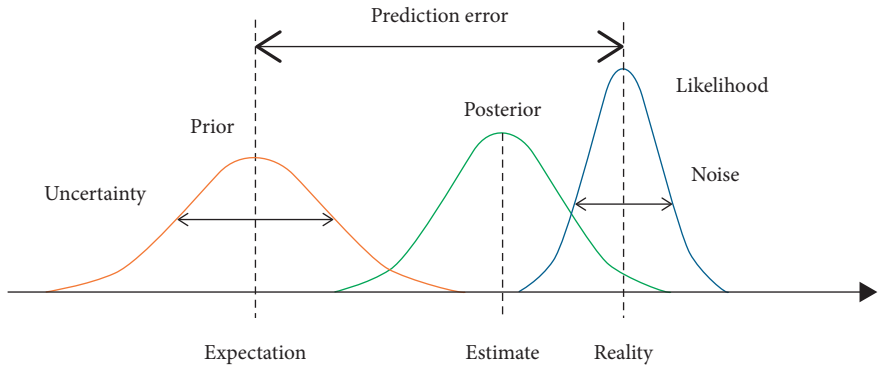

Figure 1: Bayesian inference.

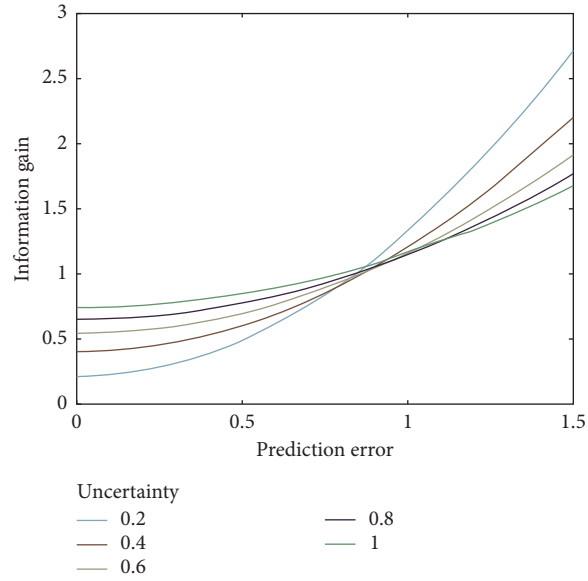

Figure 2: Information gain vs prediction error.

(2) The marginal likelihood of the data presented in the model can be used to derive a lower limit for the marginal probability of the observed data, with the marginalization conducted on unobserved variables. The main notion is that a higher marginal probability for a set of variables suggests a better fit of the data and thus a greater likelihood of a data leak in the model.

An example of information gain vs prediction error is presented in Figure 2.

Information gain is calculated mathematically as a function of prediction errors for uncertainty levels ranging from 0.2 to 1.0. The external noise level is set to 0.1 [23, 27].

The method generally approaches a conditional latent variable density given the observed variables where we assume that a mixture is present. Mixing behavior occurs because the source of each observation is unknown, that is, the classification into a specific, exact domain of a variable [32]. Thus, each observation xi is predetermined to each of $f_i(\cdot \mid \theta_i)$ with probability pi. Depending on the case, the purpose of the inference is to reconstruct the classification of observations into definition fields, construct estimators for the components' parameters, or even estimate the number of components themselves [15]. It is always feasible to map a mixture of k form distributions to a random variable Xi via a delimitation method [25, 33]:

$$\sum_{I=1}^{K} p_i f_i(x|\theta_i). \tag{1}$$

The random variable Zi with {1, 2, . . ., k}, is as follows [34]:

$$X_i|Z_i = z \sim f(x|\theta_z)\mu\varepsilon Z_i \sim M_k(1; p_1, \ldots, p_k). \tag{2}$$

Next, we assume that we have observed the extended data, which consist of independent pairs with distribution [35]:

$$P(Z_i = j|X_i = x) = \frac{p_j f_j(x)}{\sum_{I=1}^{K} p_i f_i(x)} \propto p_j f_j(x). \tag{3}$$

In the particular case of the model:

$$pN(\mu_1, 1) + (1-p)N(\mu_2, 1), \tag{4}$$

where we consider the same normal a priori distribution in the media, $\mu_1, \mu_2 \sim N(0, 10)$, we will calculate the ex post weight $\omega(z)$ for a classification z, where in the first component are l observations [24, 36]:

$$\sum_{I=1}^{N} I_{\{z_i=1\}} = l \text{ for } (n_1, n_2) = (l, n-l). \tag{5}$$

So, we have [37]

$$\pi(\underline{z}, \mu_1, \mu_2 | \underline{x}, n_1, n_2) \propto \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[I_{\{z_i=1\}}(x_i-\mu_1)^2 + \left(1-I_{\{z_i=1\}}\right)(x_i-\mu_2)^2\right] - \frac{\mu_1^2}{20} - \frac{\mu_2^2}{20}\right\} \times \frac{1}{(2\pi)^{n/2}} p^{\sum_{i=1}^{n} I_{\{z_i-1\}}} (1-p)^{n-\sum_{i=1}^{n} I_{\{z_i-1\}}}. \tag{6}$$

The ex-weight $\omega(z)$ is obtained by completing the above function in RxR for $\mu 1$ and $\mu 2$, which is a double integral which is easily calculated. For the completion in terms of $\mu 1$, excluding the parts that do not contain it, it is enough to calculate [24, 33, 36, 38]

$$I_1 = \int_{-\infty}^{+\infty} \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} I_{z_l=1}(x_i-\mu_1)^2 - \frac{\mu_1^2}{20}\right\} d\mu_1. \tag{7}$$

But

$$\begin{aligned}
&\exp\left\{-\frac{1}{2}\sum_{i=1}^{n} I_{\{z_i=1\}}(x_i-\mu_1)^2 - \frac{\mu_1^2}{20}\right\} \\
&= \exp\left\{-\frac{1}{2}\mu_1^2\left(\sum_{i=1}^{n} I_{\{z_i=1\}} + \frac{1}{10}\right) - \frac{1}{2}\sum_{i=1}^{n} x_i^2 I_{\{z_i=1\}} + \mu_1 \sum_{i=1}^{n} x_i I_{\{z_i=1\}}\right\} \\
&= \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n} I_{\{z_i=1\}} + \frac{1}{10}\right)\left(\mu_1^2 + \frac{\sum_{i=1}^{n} x_i^2 I_{\{z_i=1\}}}{\sum_{i=1}^{n} I_{\{z_i=1\}} + 1/10} - 2\mu_1 \frac{\sum_{i=1}^{n} x_i I_{\{z_i=1\}}}{\sum_{i=1}^{n} I_{\{z_i=1\}} + 1/10}\right)\right\} \\
&= \exp\left\{\frac{1}{2}\left(\sum_{i=1}^{n} I_{\{z_i=1\}} + \frac{1}{10}\right)\left(\frac{\sum_{i=1}^{n} x_i I_{\{z_i=1\}}}{\sum_{i=1}^{n} I_{\{z_i=1\}} + 1/10}\right)^2\right\} \\
&\times \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n} I_{\{z_i=1\}} + \frac{1}{10}\right)\left(\mu_1^2 + \frac{\sum_{i=1}^{n} x_i^2 I_{\{z_i=1\}}}{\sum_{i=1}^{n} I_{\{z_i=1\}} + 1/10} - 2\mu_1 \frac{\sum_{i=1}^{n} x_i I_{\{z_i=1\}}}{\sum_{i=1}^{n} I_{\{z_i=1\}} + 1/10} + \left(\frac{\sum_{i=1}^{n} x_i I_{\{z_i=1\}}}{\sum_{i=1}^{n} I_{\{z_i=1\}} + 1/10}\right)^2\right)\right\} \\
&= c_1 \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n} I_{izz,} = 1 + \frac{1}{10}\right)\left(\mu_1^2 - 2\mu_1 \frac{\sum_{i=1}^{n} x_i I_{\{z_i=1\}}}{\sum_{i=1}^{n} I_{\{z_i=1\}} + 1/10} + \left(\frac{\sum_{i=1}^{n} x_i I_{\{z_i=1\}}}{\sum_{i=1}^{n} I_{\{z_i=1\}} + 1/10}\right)^2\right)\right\} \\
&= c_1 \exp\left\{-\frac{1}{2}\left(\sum_{i=1}^{n} I_{\{z_i=1\}} + \frac{1}{10}\right)\left(\mu_1 - \frac{\sum_{i=1}^{n} x_i I_{\{z_i=1\}}}{\sum_{i=1}^{n} I_{\{z_i=1\}} + 1/10}\right)^2\right\},
\end{aligned} \tag{8}$$

$$\text{where } c_1 = \exp\left\{-\frac{1}{2}\sum_{i=1}^{n} x_i^2 I_{\{z_i=1\}} + \frac{\left(\sum_{i=1}^{n} x_i I_{\{z_i=1\}}\right)^2}{2(l+1/10)}\right\},$$

So, to calculate the integral, we have

$$I_1 = c_1 \int_{-\infty}^{+\infty} \exp\left\{ -\frac{1}{2}\left(\sum_{i=1}^{n} I_{\{z_t=1\}} + \frac{1}{10} \right)\left(\mu_1 - \frac{\sum_{i=1}^{n} x_i I_{\{z_t=1\}}}{\sum_{i=1}^{n} I_{\{z_t=\}1} + 1/10} \right)^2 \right\} d\mu_1 \Rightarrow$$
$$I_1 = c_1 \frac{\sqrt{2\pi}}{\sqrt{\sum_{i=1}^{n} I_{\{z_t=1\}} + 1/10}} = c_1 \frac{\sqrt{2\pi}}{\sqrt{l + 1/10}}. \tag{9}$$

because the last integral is crucial in the full support of the exponential distribution [39]:

$$N\left(\frac{\sum_{i=1}^{n} x_i I_{\{z_t=1\}}}{\sum_{i=1}^{n} I_{\{z_t=1\}} + 1/10}, \frac{1}{\sum_{i=1}^{n} I_{\{z_t=1\}} + 1/10} \right). \tag{10}$$

For the completion in terms of $\mu 2$, excluding the parts that do not contain it, it is enough to calculate [23, 36, 38, 40]

$$I_2 = \int_{-\infty}^{+\infty} \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} \left(1 - I_{\{z_t=1\}} \right)(x_i - \mu_2)^2 - \frac{\mu_2^2}{20} \right\} d\mu_2. \tag{11}$$

Following the same methodology as before, we conclude that [41]

$$I_1 = c_2 \frac{\sqrt{2\pi}}{\sqrt{n - l + 1/10}}$$
$$\text{where } c_2 = \exp\left\{ -\frac{1}{2} \sum_{i=1}^{n} x_i^2 \left(1 - I_{\{z_t=1\}} \right) + \frac{\left(\sum_{i=1}^{n} x_i \left(1 - I_{\{z_t=1\}} \right) \right)^2}{2(n - l + 1/10)} \right\}. \tag{12}$$

So, the ex post probability $\omega(z)$ is calculated as follows [21, 23, 42, 43]:

$$\omega(\underline{z}) = c_1 \frac{\sqrt{2\pi}}{\sqrt{l + 1/10} c_2 \sqrt{2\pi} / \sqrt{n - l + 1/10}} p^{\sum_{i=1}^{n} I_{[z-1]} - 1} (1 - p)^{n - \sum_{i=1}^{n} I_{zz} - 1\}}$$
$$= c_1 c_2 \frac{2\pi}{\sqrt{(l + 1/10)(n - l + 1/10)}} p^l (1 - p)^{n-l}. \tag{13}$$

If we replace $c1$, $c2$, we take the relation:

$$\omega(\underline{z}) = \frac{\sqrt{2\pi}}{\sqrt{(l + 1/10)}\sqrt{(n - l + 1/10)}} \times \exp\left\{ -\frac{1}{2}\left(\sum_{i=1}^{n} x_i^2 - \frac{\left(\sum_{i=1}^{n} x_i I_{\{z_t=1\}} \right)^2}{l + 1/10} - \frac{\left(\sum_{i=1}^{n} x_i \left(1 - I_{\{z_t=1\}} \right) \right)^2}{n - l + 1/10)} \right) \right\} p^l (1 - p)^{n-l}. \tag{14}$$

Thus, from the above analysis, it appears that it is practically possible to arrive at detailed expressions of the maximum probability and Bayes estimators [44] for the ex ante distributions of the variables of interest and thus marginalize the set of variables for models where there is a data leak [28, 33].

3. Experiments and Results

A specialized scenario was implemented to model the proposed system that uses sports wearables data to record the movements of athletes playing beach volleyball. The dataset comprises three-dimensional acceleration data from joint actions of beach volleyball athletes, each of whom was fitted with an accelerometer worn on the wrist and sampled at 39 Hz. The signal was recorded at 14 bits per axis and then compressed to 16 g. The x, y, and z axes relate to the athletes' spatial arrangement, which is recorded in an independent coordinate system based on the sensor configuration, as there was no transfer to real-world coordinates [45, 46]. The 30 athletes recorded ranged in expertise from novice to professional

volleyball players. The set's goal is to create an identification and classification system that extracts relevant portions from continuous input and classifies them [47]. The categorization includes ten various volleyball activities, such as homemade service, block, nail, and so on. For the evaluation of the system, 10 characteristics were selected, which were randomly combined into pairs to identify the observed variables, whether they come from a coupling method and whether there is a data leak.

We first describe some key features. Let $g(\cdot,\cdot|\theta)$ be the joint density function of (X, Z) given by the parametric vector θ, $f(\cdot|\theta)$ be the density function of X given θ, and $k(\cdot|x,\theta)$ be the function density of the bounded distribution of Z given by observations x and θ. The algorithm is based on the use of incomplete data, i.e., we can write the distribution of sample x as follows [1, 2, 40]:

$$\begin{aligned} f(x|\theta) &= \int g(\underline{x},\underline{z}\,|\theta)\mathrm{d}\underline{z} \\ &= \int f(x|\theta)k(\underline{z}\,|\,\underline{x},\theta)\mathrm{d}\underline{z}. \end{aligned} \tag{15}$$

So, logarithm it:

$$g(x,z|\theta) = f(x|\theta)k(\underline{z}\,|\,\underline{x},\theta). \tag{16}$$

We arrive at a complete (unobserved) logarithm of probability:

$$L^c(\theta|\,\underline{x},\underline{z}) = L(\theta|\,\underline{x}) + \log k(\underline{z}\,|\,\underline{x},\theta), \tag{17}$$

where L is the observed logarithm of the probability. The algorithm fills in the missing variables z based on $k(z|x, \theta)$ and then maximizes with θ the expected full logarithm probability [21, 25, 48].

So, the algorithm is configured as follows:

(1) Give some initial values to $\theta(0)$.

(2) For each t, $t = 1, 2, \ldots, n$, calculate $Q(\theta|\theta^{(t-1)}, \underline{x}) = E_{\theta^{(t-1)}}(L^c(\theta|\,\underline{x},\underline{Z}))$ where $\underline{Z} \sim k(z|x,\theta)$.

(3) Maximize concerning θ the $Q(\theta|\theta^{(t-1)}, \underline{x})$ and set $\theta^{(t)} = \arg\max_{\theta} Q(\theta|\theta^{(t-1)}, \underline{x})$.

When performing the above algorithm, the result is that in each iteration, the (observed) $L(\theta|x)$ increases.

As an application of the above, we consider the particular case of the model of mixing two regular variables, where all parameters are known except $\theta = (\mu 1, \mu 2)$. For a simulated sample of 500 observations and actual values $p = 0.7$ and $(\mu 1, \mu 2) = (0, 2.5)$, the logarithm of probability has two peaks. Applying the algorithm to this model, we have that the total probability is [20, 49, 50]

$$p^{\sum_{i=1}^{n} I_{izi}-1}(1-p)^{n-\sum_{i=1}^{n} Iz_i-1\}}(2\pi)^{-n/2} \cdot \exp\left\{-\frac{1}{2}\sum_{i=1}^{n}\left[I_{\{z_t=1\}}(x_i-\mu_1)^2+\left(1-I_{\{z_t=1\}}\right)(x_i-\mu_2)^2\right]\right\}, \tag{18}$$

where its logarithm is

$$\begin{aligned} L^c(\theta\underline{x},\underline{z}) &= \sum_{i=1}^{n} I_{\{z_t=1\}}\log p + \left(n-\sum_{i=1}^{n} I_{\{z_t=1\}}\right)\log(1-p) \\ &\quad -\frac{n}{2}\log(2\pi) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\left[I_{\{z_t=1\}}(x_i-\mu_1)^2+\left(1-I_{\{z_t=1\}}\right)(x_i-\mu_2)^2\right]. \end{aligned} \tag{19}$$

For the first step, we need to calculate

$$Q\left(\theta|\theta^{(t-1)},\underline{x}\right) = E_{\theta^{(t-1)}}(\log L^c(\theta|\,\underline{x},\underline{Z})), \tag{20}$$

where the mean value is taken for $\underline{Z} \sim k(z|x,\theta)$, and we have that Zi are independent of [51–54]

$$\begin{aligned} &P(Z_i = 1|\,\underline{\theta},\underline{x}) \\ &= \frac{p\exp\left\{-(x_i-\mu_1)^2/2\right\}}{p\exp\left\{-(x_i-\mu_1)^2/2\right\}+(1-p)\exp\left\{-(x_i-\mu_2)^2/2\right\}} \\ &= 1 - P(Z_i = 2 \mid \underline{\theta},\underline{x}). \end{aligned} \tag{21}$$

In step t, the expected rankings are equal to

$$\begin{aligned} \hat{z}_i^{(t-1)} &= E\left(\sum_{i=1}^{n} I_{\{z_t=1\}}|\overset{(t-1)}{\underline{\theta}},\underline{x}\right) \\ &= P\left(Z_i = 1|\overset{(t-1)}{\underline{\theta}},\underline{x}\right). \end{aligned} \tag{22}$$

Therefore:

$$\begin{aligned} Q\left(\theta|\theta^{(t-1)},\underline{x}\right) &= \sum_{i=1}^{n}\hat{z}_i^{(t-1)}\log p + \left(n-\sum_{i=1}^{n}\hat{z}_i^{(t-1)}\right)\log(1-p) \\ &\quad -\frac{n}{2}\log(2\pi) \\ &\quad -\frac{1}{2}\sum_{i=1}^{n}\left[\hat{z}_i^{(t-1)}(x_i-\mu_1)^2+\left(1-\hat{z}_i^{(t-1)}\right)(x_i-\mu_2)^2\right]. \end{aligned} \tag{23}$$

which we maximize in the second step in terms of $(\mu 1, \mu 2)$ and get

$$\begin{aligned} \mu_1^{(t)} &= \frac{\sum_{i=1}^{n}\hat{z}_i^{(t-1)}x_i}{\sum_{i=1}^{n}\hat{z}_i^{(t-1)}}, \\ \mu_2^{(t)} &= \frac{\sum_{i=1}^{n}1-\hat{z}_i^{(t-1)}x_i}{\sum_{i=1}^{n}1-\hat{z}_i^{(t-1)}}. \end{aligned} \tag{24}$$

This example involved running the algorithm 20 times (each time with 100 repeats) while picking random numbers from a range of possibilities for the initial conditions. However, the proposed approach was only drawn to the highest and principal vertex of the logarithm

probability eight times out of every 20 times in the experiments. It was drawn to the pseudo-vertex of the logarithm probability distribution for the remaining 12 times (although the likelihood is much lower). The original values were closer to the lower peak than the final values, indicating that the early values were more accurate. The algorithm converges to the pseudo-peak of likelihood, at which point we may make 84 percent correct predictions about the coupling between the variables in the dataset. Accordingly, we will have 93 percent of the variables accurately predicted to couple their coefficients if the algorithm converges to the dominant peak in probability.

4. Discussion and Conclusions

In this work, we proposed an innovative system of leakage prediction in machine learning models, which is based on Bayesian inference, to calculate a lower limit for the marginal probability of the observed variables coming from a coupling method, which shows that in an examined machine learning model, there is data leakage. The methodology is evaluated in a specialized dataset from sports wearable sensors, where the ability of the method to detect variable coupling is demonstrated, even when it is done randomly.

The proposed methodology is a Bayesian approach to statistical discoveries in complicated distributions that are difficult to evaluate directly or by sampling, and this is the methodology that has been offered. It is a method of selection that is different from Monte Carlo sampling methods. While Monte Carlo techniques use a sequence of samples to approximate a rear distribution numerically, the proposed algorithm provides a locally optimal, correct analytical solution, allowing even hidden variable coupling to be found. From the maximum ex post estimate of each variable's unique most probable value to the fully Bayesian estimation that calculates (approximately) the entire rear distribution of parameters and latent variables, the algorithm finds a set of optimal parameters of the interrelated variables, which can then be solved in detail using the information obtained from the data. Indeed, this is true even for conceptually comparable variables, such as a basic nonhierarchical model with only two parameters and no latent variables.

The extension of the methodology can focus on integrating countervailing machine learning techniques to be a complete defense system in case of attacks that attempt to deceive the models by providing misleading information. Determine strategies and procedures for running the model on specified sets of issues with training and test data generated from the same statistical distribution. Moreover, a future expansion of the proposed system will review the taxonomies of the characteristics of transfer learning, particularly whether and how this system can mitigate them. Finally, learning transfer approaches are investigated from known distribution attack methods seeking to exploit the dynamics of categorization decision-making limits.

Data Availability

The data used in this study are available from the corresponding author upon request.

Conflicts of Interest

The author declares that there are no conflicts of interest.

References

[1] K. M. R. Alam, N. Siddique, and H. Adeli, "A dynamic ensemble learning algorithm for neural networks," *Neural Computing & Applications*, vol. 32, no. 12, pp. 8675–8690, 2020.

[2] J. Gawlikowski, "A Survey of Uncertainty in Deep Neural Networks," 2021, http://arxiv.org/abs/2107.03342.

[3] K. Demertzis, L. Iliadis, and P. Kikiras, "A Lipschitz - Shapley Explainable Defense Methodology against Adversarial Attacks," in *Proceedings of the Artificial Intelligence Applications and Innovations. AIAI 2021 IFIP WG 12.5*, pp. 211–227, Crete, Greece, June, 2021.

[4] R. Chauhan and S. Shah Heydari, "Polymorphic Adversarial DDoS attack on IDS using GAN," in *Proceedings of the 2020 International Symposium on Networks, Computers and Communications (ISNCC)*, pp. 1–6, Shenzhen, China, July. 2020.

[5] Q. Liu, J. Guo, C.-K. Wen, and S. Jin, "Adversarial attack on DL-based massive MIMO CSI feedback," *Journal of Communications and Networks*, vol. 22, no. 3, pp. 230–235, 2020.

[6] P. Yu, K. Song, and J. Lu, "Generating adversarial examples with conditional generative adversarial net," in *Proceedings of the 2018 24th International Conference on Pattern Recognition (ICPR)*, pp. 676–681, Beijing, China, August. 2018.

[7] Z.-A. Zhu, Y.-Z. Lu, and C.-K. Chiang, "Generating adversarial examples by makeup attacks on face recognition," in *Proceedings of the 2019 IEEE International Conference on Image Processing (ICIP)*, pp. 2516–2520, Taipei, Taiwan, September. 2019.

[8] J. Yu, Y. Lee, K. C. Yow, M. Jeon, and W. Pedrycz, "Abnormal event detection and localization via adversarial event prediction," *IEEE Transactions on Neural Networks and Learning Systems*, no. –15, pp. 1–15, 2021.

[9] Z. Shi, Y. Ma, and X. Yu, "An effective and efficient method for word-level textual adversarial attack," in *Proceedings of the 2021 IEEE Symposium on Computers and Communications (ISCC)*, pp. 1–6, Athens, Greece, September. 2021.

[10] P. Tang, W. Wang, J. Lou, and L. Xiong, "Generating adversarial examples with distance constrained adversarial imitation networks," *IEEE Transactions on Dependable and Secure Computing*, p. 1, 2021.

[11] B. Tarchoun, I. Alouani, A. Ben Khalifa, and M. A. Mahjoub, "Adversarial attacks in a multi-view setting: an empirical study of the adversarial patches inter-view transferability," in *Proceedings of the 2021 International Conference on Cyberworlds (CW)*, pp. 299–302, Caen, France, September. 2021.

[12] P. Gattineni and G. S. Dharan, "Intrusion Detection Mechanisms: SVM, random forest, and extreme learning machine (ELM)," in *Proceedings of the 2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pp. 273–276, Coimbatore, India, September. 2021.

[13] P. Rathore, A. Basak, S. H. Nistala, and V. Runkana, "Untargeted, targeted and universal adversarial attacks and defenses on time series," in *Proceedings of the 2020*

International Joint Conference on Neural Networks (IJCNN), pp. 1–8, Glasgow, UK, July. 2020.

[14] A. Salem, Y. Zhang, M. Humbert et al., "Model and Data Independent Membership Inference Attacks and Defenses on Machine Learning Models," 2018, http://arxiv.org/abs/1806.01246.

[15] Y. Lu, X. Huang, Y. Ma, and M. Ma, "A weighted context graph model for fast data leak detection," in *Proceedings of the 2018 IEEE International Conference on Communications (ICC)*, pp. 1–6, Kansas City, MO, USA, May, 2018.

[16] M. Miyatake, H. Sawai, Y. Minami, and K. Shikano, "Integrated training for spotting Japanese phonemes using large phonemic time-delay neural networks," *International Conference on Acoustics, Speech, and Signal Processing*, vol. 1, pp. 449–452, 1990.

[17] J. O. Berger, "Bayesian analysis, Springer Series in Statistics," in *Statistical Decision Theory and Bayesian Analysis*, pp. 118–307, Springer, New York, NY, USA, 1985.

[18] J. O. Berger, "Basic concepts," in *Statistical Decision Theory and Bayesian Analysis*, pp. 1–45, Springer, New York, NY, USA, 1985.

[19] D. P. Kingma and M. Welling, "Auto-Encoding Variational Bayes," 2014, http://arxiv.org/abs/1312.6114.

[20] A. J. M. Garrett, "Review: probability theory: the logic of science," *Probability and Risk*, vol. 3, no. 3-4, pp. 243–246, 2004.

[21] L. E. B. Salasar, J. G. Leite, and F. Louzada, "Likelihood-based inference for population size in a capture-recapture experiment with varying probabilities from occasion to occasion," *Brazilian Journal of Probability and Statistics*, vol. 30, no. 1, pp. 47–69, 2016.

[22] J. Lü and P. Wang, "Modeling and analysis of large-scale networks," in *Modeling and Analysis of Bio-Molecular Networks*, pp. 249–292, Springer, Singapore, 2020.

[23] Y. Emma Wang, Y. Zhu, G. G. Ko, B. Reagen, G.-Y. Wei, and D. Brooks, "Demystifying bayesian inference workloads," in *Proceedings of the 2019 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, pp. 177–189, Madison, WI, USA, March, 2019.

[24] S. Jun, "Bayesian Inference and Learning for Neural Networks and Deep Learning," in *Proceedings of the 2020 International Conference on Artificial Intelligence in Information and Communication (ICAIIC)*, pp. 569–571, Seoul, Korea, Octobar. 2020.

[25] Z. Rudong, S. Xianming, W. Qian, S. Xiaobo, and S. Xing, "Bayesian inference for ammunition demand based on Gompertz distribution," *Journal of Systems Engineering and Electronics*, vol. 31, no. 3, pp. 567–577, 2020.

[26] Handbook of Statistics, "Bayesian Thinking, Modeling and Computation - PDF Free Download," 2022, https://epdf.tips/handbook-of-statistics-volume-25-bayesian-thinking-modeling-and-computation.html.

[27] H. Yanagisawa, O. Kawamata, and K. Ueda, "Modeling emotions associated with novelty at variable uncertainty levels: a bayesian approach," vol. 13, 2019 https://www.frontiersin.org/article/10.3389/fncom.2019.00002.

[28] X.-d. Zhang, "An improved bayesian network inference algorithm," in *Proceedings of the 2010 Third International Conference on Intelligent Networks and Intelligent Systems*, pp. 389–392, Shenyang, China, August. 2010.

[29] J. Yun-Jie, C. Wen-Qi, and H. Ling, "Risk identification and simulation based on the bayesian inference," in *Proceedings of the 2018 4th Annual International Conference on Network and Information Systems for Computers (ICNISC)*, pp. 407–411, Wuhan, China, April. 2018.

[30] D. Hou, T. Driessen, and H. Sun, "The Shapley value and the nucleolus of service cost savings games as an application of 1-convexity," *IMA Journal of Applied Mathematics*, vol. 80, no. 6, pp. 1799–1807, 2015.

[31] G. Alessandrini, M. V. D. Hoop, R. Gaburro, and E. Sincich, "Lipschitz stability for a piecewise linear Schrödinger potential from local Cauchy data," *Asymptot. Anal*, vol. 108, no. 3, pp. 115–149, 2018.

[32] "Permutation principles for the change analysis of stochastic processes under strong invariance," 2022, https://dl.acm.org/doi/abs/10.5555/1124448.1716910.

[33] J. Barbier, "Overlap matrix concentration in optimal Bayesian inference," *Information and Inference: A Journal of the IMA*, vol. 10, no. 2, pp. 597–623, 2020.

[34] O. Lee, "Probabilistic properties of a nonlinear ARMA process with markov switching," *Communications in Statistics - Theory and Methods*, vol. 34, no. 1, pp. 193–204, 2005.

[35] Y. Lu, X. Huang, D. Li, and Y. Zhang, "Collaborative graph-based mechanism for distributed big data leakage prevention," in *Proceedings of the 2018 IEEE Global Communications Conference (GLOBECOM)*, pp. 1–7, Abu Dhabi, UAE, September. 2018.

[36] M. T. Koudahl and B. de Vries, "Batman: bayesian target modelling for active inference," in *Proceedings of the 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 3852–3856, Barcelona, Spain, Febrauary. 2020.

[37] D. K. Dussmann, "Computational Systems Biology," 2022, https://www.kulturkaufhaus.de/en/detail/ISBN-2244012260139/Lecca-Paola/Computational-Systems-Biology.

[38] E.-H. Choi, T. Fujiwara, and O. Mizuno, "Weighting for combinatorial testing by bayesian inference," in *Proceedings of the 2017 IEEE International Conference on Software Testing, Verification and Validation Workshops (ICSTW)*, pp. 389–391, Tokyo, Japan, March, 2017.

[39] S. Fan, Y. Wang, and L. Xiao, "Multidimensional BSDEs with uniformly continuous generators and general time intervals," *Bulletin of the Korean Mathematical Society*, vol. 52, no. 2, pp. 483–504, 2015.

[40] Z. Fei, K. Liu, B. Huang, Y. Zheng, and X. Xiang, "Dirichlet process mixture model based nonparametric bayesian modeling and variational inference," in *Proceedings of the 2019 Chinese Automation Congress (CAC)*, pp. 3048–3051, Hangzhou, China, August. 2019.

[41] X. Hong, "Study of intergenerational mobility and urbanization based on OLS method and ordered probit mode," in *Proceedings of the 2020 Management Science Informatization and Economic Innovation Development Conference (MSIEID)*, pp. 435–447, Guangzhou, China, September. 2020.

[42] H. Chen and J. Ren, "Structure-variable hybrid dynamic bayesian networks and its inference algorithm," in *Proceedings of the 2012 24th Chinese Control and Decision Conference (CCDC)*, pp. 2815–2820, Taiyuan, China, Febrauary. 2012.

[43] H. Guan, J.-C. Ni, Q. Zhang, L. Sun, and K. Wang, "Saliency detection for $\mathbf{L}_{1/2}$ regularization-based SAR image feature enhancement via bayesian inference," in *Proceedings of the IGARSS 2018 - 2018 IEEE International Geoscience and Remote Sensing Symposium*, pp. 4483–4486, Valencia, Spain, July. 2018.

[44] Z. Lijun, H. Guiqiu, L. Qingsheng, and D. Guanhua, "An intuitionistic calculus to complex abnormal event recognition

on data streams," *Security and Communication Networks*, vol. 2021, pp. 1–14, 2021.

[45] T. Kautz, B. H. Groh, J. Hannink, U. Jensen, H. Strubberg, and B. M. Eskofier, "Activity recognition in beach volleyball using a deep convolutional neural network," *Data Mining and Knowledge Discovery*, vol. 31, no. 6, pp. 1678–1705, 2017.

[46] J. Link, T. Perst, M. Stoeve, and B. M. Eskofier, "Wearable sensors for activity recognition in ultimate frisbee using convolutional neural networks and transfer learning," *Sensors*, vol. 22, no. 7, p. 2560, 2022.

[47] T. Aira, K. Salin, T. Vasankari et al., "Training volume and intensity of physical activity among young athletes: the health promoting sports club (HPSC) study," *Advances in Physical Education*, vol. 09, no. 04, pp. 270–287, 2019.

[48] H. Worthington, R. S. McCrea, R. King, and R. A. Griffiths, "Estimation of population size when capture probability depends on individual states," *Journal of Agricultural, Biological, and Environmental Statistics*, vol. 24, no. 1, pp. 154–172, 2019.

[49] M. Burgin and P. Rocchi, "Ample probability in cognition," in *Proceedings of the 2019 IEEE 18th International Conference on Cognitive Informatics & Cognitive Computing (ICCI∗CC)*, pp. 62–65, Milan, Italy, July. 2019.

[50] S. Guopan, "The effect of probability on risk perception and risk preference in decision making," in *Proceedings of the 2010 International Conference on Education and Management Technology*, pp. 690–693, Wasinghton, USA, November. 2010.

[51] T. M. F. Alves, R. O. J. Soeiro, and A. V. T. Cartaxo, "Probability distribution of intercore crosstalk in weakly coupled MCFs with multiple interferers," in *Proceedings of the 2019 IEEE Photonics Conference (IPC)*, pp. 1–4, San Antonio, TX, USA, September. 2019.

[52] B. H. H. Gade, C. N. Vooren, and M. Kloster, "Probability distribution for association of maneuvering vehicles," in *Proceedings of the 2019 22th International Conference on Information Fusion (FUSION)*, pp. 1–7, Ottawa, Canada, July. 2019.

[53] H. Igarashi and K. Watanabe, "Complex adjoint variable method for finite-element analysis of eddy current problems," *IEEE Transactions on Magnetics*, vol. 46, no. 8, pp. 2739–2742, 2010.

[54] J. Qian, J. P. Lu, S. L. Hui, Y. J. Ma, and D. Y. Li, "Dynamic analysis and CFD numerical simulation on backpressure filling system," *Mathematical Problems in Engineering*, vol. 2015, Article ID 160641, 8 pages, 2015.