AIMS Mathematics, 11(4): 11731–11759.
DOI: 10.3934/math.2026483
Received: 11 December 2025
Revised: 12 March 2026
Accepted: 03 April 2026
Published: 28 April 2026

https://www.aimspress.com/journal/Math

Research article

Markov processes for enhanced deepfake generation and detection

Michael A. Kouritzin[1], **Ian Zhang**[2,*], **Jyoti Bhadana**[3] **and Seoyeon Park**[1]

[1] Department of Mathematical and Statistical Sciences, University of Alberta, Edmonton, AB, Canada

[2] Department of Statistical Sciences, University of Toronto, Toronto, ON, Canada

[3] Department of Mathematics, University of Texas at Arlington, Arlington, TX, USA

* **Correspondence:** Email: ianz.zhang@mail.utoronto.ca; Tel: +17809092566.

Abstract: We investigate both new and existing methods for generating, and especially detecting, deepfakes through the simple but informative task of authenticating binary coin flip data. The main contribution is the introduction of a Markov observation model (MOM) as an alternative probabilistic framework for both deepfake generation and discrimination. Its performance is compared against several existing approaches, such as generative adversarial networks (GANs), support vector machines (SVMs), branching particle filtering (BPF), and human alternatives. Since SVMs are discriminative methods and do not have generative abilities, they are only evaluated for the detection task, while the remaining approaches are assessed on both generation and discrimination. Across the experiments, human participants are shown to perform the worst, which demonstrates the difficulty of reliably identifying deepfaked sequences by a human eye. Among the computational methods, GANs perform better than humans, but are outperformed by SVMs, which in turn are surpassed by BPF. The strongest overall performance comes from the proposed MOM approach, which achieves the best results for deepfake detection out of all methods. A similar result is observed for the generation task, with MOM again showing the strongest performance, followed by BPF, GAN, and humans. These results showcase the generative and discrimination abilities of the proposed method.

Keywords: deepfakes; generative adversarial networks; Markov chain; detection; likelihood; model selection; simulation
Mathematics Subject Classification: 62M05; 60J22; 65C40

1. Introduction

Fake or synthetic content generated though advanced deep learning techniques that appears authentic in the eyes of a human being is called a “deepfake”. The most common form of deepfakes

involves the generation and manipulation of human imagery. Deepfake technology has creative and productive applications in entertainment, education, content creation, computer vision, natural language processing, and human-level control [32]. Deepfakes have spread to other domains and media such as forensics, finance, and healthcare [29]. However, deepfakes also pose substantial risks such as misinformation, privacy invasion, and identity theft. While they are usually trained to foil discriminators, the objective of the generated content is to fool a human, not a machine, and for that reason they have garnered heightened attention. Still, a machine can also be a valuable tool in its detection.

Rather than studying full audio-visual deepfakes directly, this paper focuses on a controlled coin-flip sequence setting. This lets us compare different detection and generation methods under clear data-generating assumptions, while still keeping the core fake-versus-real classification problem.

This problem is related to earlier work on distinguishing truly random sequences from fake sequences produced by humans or machines [13, 17, 31]. It is also connected to probabilistic sequence models such as hidden Markov models and pairwise Markov chains, which provide natural tools for sequence generation, filtering, and likelihood-based classification [23, 24].

We compare several approaches in this controlled sequence setting, including GAN, SVM, branching particle filtering (BPF), and the proposed Markov observation model (MOM). MOM replaces the neural-network generator/discriminator framework with a latent probabilistic sequence model that supports both sequence generation and likelihood-based classification. This allows us to study the trade-off between probabilistic structure, detection performance, and computational cost in a setting where the assumptions are explicit and the methods are directly comparable.

There is a long-standing statistical perspective for analyzing sequential data through probabilistic sequence models, including Markov-chain and hidden Markov model (HMM) frameworks. These models support likelihood-based inference, recursive decoding/inference algorithms, and interpretable parametrization, and they are often studied together with model-selection tools such as information criteria and likelihood-ratio based procedures [5, 24].

Many researchers have explored the classification of coin-flip sequences as real or fake [26, 28], whereas others have developed algorithms for generating fake flips whilst capturing the properties of a real flip [17]. Another example of this problem proposed in [3] was addressed to school aged children, with the task being that they needed to correctly identify which sequences were "real" or "random", along with which sequences appeared to be "produced by a fair coin" or "fake" [25]. This problem was then used to understand and model the "fake" behavior. The power in the existing probabilistic approach comes from a separate hidden model, the signal model, that represents both the real and fake random behaviors, while the variables of interest, in this instance the coin flips, are modeled as the observations [17] that depend upon this signal. Generation then comes from signal, observation simulation. Detection, on the other hand, becomes estimating which behavior is present in the signal, which can be done optimally by filtering theory [33] and practically by particle filters [7, 8, 12, 21]. This existing approach entails encoding all fake and real models into the same signal, usually by including some state variable that indicates which behavior is active. The problems with expanding this approach beyond simple authentication problems like coin flips is that one needs to: i) have models for every type of fake as well as real behavior, ii) amalgamate these models together into one signal model and iii) suffer extra processing costs by working with all behaviors simultaneously when they are amalgamated into one signal. This does not seem practical for large problems like authenticating

real video content. We will suggest a new probabilistic approach below that both outperforms this existing approach and is more expandable to larger problems. The first step of the new approach is separating out the signal and employing recursive likelihood model selection, which we will also use herein with branching particle filters (BPFs). The recursive likelihood approach was introduced in [16]. BPFs, where branching is used in place of resampling to redistribute particles, were introduced in [6], but these early algorithms suffered from extreme particle swings, which affected performance dramatically. [1] developed a stabler BPF. However, it was only in [14, 15] that a fully stable algorithm was given where it could be shown that the number of particles in a BPF was controllable. It was also shown there that BPFs can have recursive likelihood model selection capabilities with effective resampling. Finally, it was also shown that by reducing particle number fluctuations, one improves performance and reliability [15].

To address these limitations, we introduce a new GAN-like architecture that utilizes a model and algorithm recently proposed by [16]. In particular, fake sequences are modeled as Markov chains with hidden states. This Markov observation model (MOM) replaces the neural networks within the traditional GAN architecture. MOM is simply a pairwise Markov chain $(X, Y)'$ with one step transition probabilities p and initial distribution μ, where one component of the chain X is hidden and the other Y is observable. It was shown in [16] that a Baum-Welch-like forward-backward expectation-maximization (EM) algorithm for both p and μ holds in this setting and converges to (at least) local maxima of the likelihood function. Like for HMM, the forward-backward reduction of the computations produces a highly efficient learning algorithm. The *generation* part of MOM comes, after the learning is complete, by simulating the pairwise Markov chain and then throwing the hidden part away. The *discrimination* part of MOM comes, after learning all possible (real and fake) models, by likelihood model selection, which is also shown in [16], to determine if a given sequence is better represented by real or fake models.

We chose to introduce our method on the simple coin flip problem in order to compare it to many other methods and due to the availability of computer resources. However, there is no need to encode mathematical models into a signal, and there are efficient forward-backward algorithms for learning as well as classifying. Hence, there appears to be no reason why the MOM-based method cannot be expanded to more practical and pressing deepfake detection problems in the future.

Our work is also related to broader research on robust multimodal learning, including work on multimodal fusion, modality balancing, robustness to missing or noisy modalities, and lightweight or privacy-aware multimodal methods [2, 34, 35]. More broadly, much of the deepfake detection literature focuses on artifact cues, deep neural architectures for image and video data, and reliability challenges such as transferability and robustness [19, 27, 30]. These papers show the broader importance of reliable discrimination in difficult settings, but they study much richer multimodal problems than the one considered here. Our paper instead studies a controlled binary-sequence benchmark, which we use as a simple and interpretable test setting rather than a full multimodal audio-visual deepfake benchmark [20, 27].

1.1. Layout

This paper is mainly divided into two parts: one is deepfake generation, and another is deepfake detection. Specifically, in Section 2, we discuss methods of simulating a fake coin flip sequence. In Section 3, we discuss methods of classification. In Section 4, we discuss adversarial tuning, and in

Section 5 we discuss our experimentation and comparative results. Discussion and conclusions are given in Section 6.

1.2. Computer system

All simulations were done on an *M3 Pro MacBook Pro* in Python 3.10. In our study, we utilized a variety of packages and libraries to conduct simulations and analyses. The core libraries included NumPy for numerical computations, Pandas for data handling and analysis, and the random module for generating random numbers. We employed the Keras library to build and train the GAN model, leveraging layers such as Dense, LeakyReLU, BatchNormalization, and optimizers like Adam. We used scikit-learn, implementing models like support vector classification (SVC) and one-class SVM, and performed tasks such as cross-validation, and train-test splitting. Additionally, we assessed model performance using metrics from scikit-learn, including precision, recall, classification reports, ROC-AUC, average precision scores, and confusion matrices. These tools were integral to processing data, building models, and evaluating outcomes within our simulations.

2. Methods of data generation

This section describes the different ways coin flip sequences were created. We will use five sources: real (by random number generator), human fake, the simulator algorithm that can be thought of as the generator part of the BPF approach, the generative part of GAN, and the generative part of MOM. Generally, some initial data is created representing the two types, real and (deep) fake, to distinguish. Then, adversarial methods are used to create more deepfake data. Some initial deepfake data was produced by a Bernoulli generation algorithm from [17], termed the *simulator* here, capable of producing desired correlations to prior samples. Other deepfake data was created using generative adversarial networks (GAN), and the newly proposed Markov observation model (MOM) generation.

2.1. Initial data

Adversarial networks require some initial training data or else some other means of setting up initial generator and discriminator rules. We started the system with real, fake, and deepfake data sequences as follows.

2.1.1. Real sequences

We used Python's `random` package on an M3 Pro MacBook Pro to generate independent sequences of independent coin flip data. We treated these as if they were real coin flips, even though they were computer generated. For representation purposes, ones were interpreted as heads, and zeroes were interpreted as tails.

2.1.2. Real fakes

We employed a group 15 students to create 137 sequences of 200 *real fake* coin flips that they thought would fool other students into thinking they were real sequences. These students had at least a basic background in probability and were aware of some of the artifacts that were likely in real sequences.

2.1.3. Initial deepfakes, the filtering generator

[17] used a method of simulating coin-flip sequences with prescribed pairwise covariances (between flips at two points in the sequence) and marginal probabilities (of heads or tails), which we will call the *simulator*. The simulator can, for example, produce a fair sequence where neighbor flips are slightly negatively (positively) correlated and others further away are positively (negatively) correlated to compensate. Here, the authors reduce the computational requirements of the coin flip simulations by using an explicit formula that encodes the desired covariances. Actually, this method turned out to be extendable to discrete (not just binary) random variables on graphs and to be quite useful in applications (see [18, 22]).

We describe the process of simulating deepfakes using the simulator. Let r_k represent the probability of getting a 1 on the k^{th} flip Y_k and $\beta^l_{k,j}$ be the covariance $\beta^l_{k,j} = \text{cov}(Y_k, Y_{k-j})$. Then, the model for the trivial faker is:

$$\text{(TF)} \quad r_{k+1} = r_k + \varepsilon r_k(1 - r_k)\xi^P_k \quad \text{and} \quad \beta^l_{k+1,j} = \beta^l_{k,j} + \varepsilon\beta^l_{k,j}(\beta^l_{k,j} + 1)(1 - \beta^l_{k,j})\xi^j_k$$

for $k = 0, ..., 199$, $j = 1, ..., l$. Here, $\{\xi^P_k\}_{k=1}^{\infty}$ and $\{\xi^j_k\}_{j,k=1}^{l,\infty}$ are $p = \frac{1}{2}$, $\{-1, 1\}$-Bernoulli independent of everything and ε is a small parameter. Fakers that try to undo what has been done follow the random sign change model

$$\text{(RSC)} \quad r_{k+1} = r_k + \rho^P_k(1 - 2r_k) + \varepsilon r_k(1 - r_k)\xi^P_k \quad \text{and} \quad \beta^l_{k+1,j} = \rho_{k,j}\beta^l_{k,j} + \varepsilon\beta^l_{k,j}(\beta^l_{k,j} + 1)(1 - \beta^l_{k,j})\xi^j_k$$

for $k = 0, ..., 199$, $j = 1, ..., l$. Here, $\{\rho^P_k\}$ and $\{\rho_{k,j}\}$ are independent such that

$$P(\rho_{k,j} = -1) = P(\rho^P_k = 1) = 1 - P(\rho_{k,j} = 1) = 1 - P(\rho^P_k = 1) = \delta$$

for some small $\delta > 0$. The trivial faker and the random sign change faker are our initial deepfake models. Note that both are designed to keep $r_k \in [0, 1]$ and $\beta^l_{k,j} \in [-1, 1]$ if started that way. They must be supplied with initial values $r_0 \approx \frac{1}{2}$ (for fairness) and $\beta_{k,j} \approx 0$ for $k - j \leq 0$ (for near independence).

The real coin flips can then be modeled in this way and fit into the algorithm

$$\text{(Real)} \quad r_k = \frac{1}{2} \quad \text{and} \quad \beta^l_{k,j} = 0$$

for $k = 1, ..., 200$, $j = 1, ..., l$. Then, for the signal we set

$$\Theta = \begin{cases} 1 & \text{Trivial Faker} \\ 0 & \text{RSC Faker} \\ -1 & \text{Real Coin} \end{cases}$$

and let $X_k = \begin{bmatrix} \Theta & r_k & \{\beta^l_{k,j}\}^l_{j=1} \end{bmatrix}'$ be the signal. The goal of filtering is to estimate X_k optimally from the back observations $Y_l, l \leq k$, which includes the goal of deepfake detection. The goal of deepfake detection is to determine the value of Θ given the observations.

For this method, we initially simulated 137 sequences of 200 coin flips with the above equation and algorithm, which were deemed to be deepfakes due to the known difficulty in distinguishing them from real as well as their computer algorithm generation.

The algorithm is designed to simulate sequences of coin flips that align with specified marginal probabilities and pairwise covariances. We begin the simulation by defining the total number of flips (N_f) and the number of pairwise covariances (N_c) and the number of pairwise covariances. These parameters dictate the length of the sequence and the complexity of the dependencies between flips. These parameters should be uniformly distributed within specified ranges to ensure variability across trials. Apply the algorithm to simulate a sequence of coin flips. For each flip, compute the probability of it being heads based on the outcomes of the previous flips, then use the uniform random variable U to determine the flip's result. After generating the sequence, estimate the marginal probability $\bar{r}$ and the pairwise covariances $\{\bar{\beta}_j^{Nc}\}$ from the simulated data. These estimates will be used to evaluate how closely the simulated sequence matches the desired statistical properties. Calculate the error for the trial using the formula, $\text{err} = \frac{(r-\bar{r})^2+\sum_{j=1}^{Nc}(\beta_j^{Nc}-\bar{\beta}_j^{Nc})^2}{N_c+1}$. This error metric combines the discrepancies in both the marginal probability and the pairwise covariances, normalized by the number of covariances plus one. By averaging the error over multiple trials, one can assess the algorithm's performance and its ability to generate sequences that accurately reflect the specified probabilities and covariances. The reader is referred to the work [18] for further applied properties and a more detailed explanation of the algorithm.

2.2. One step learning

After the initial real, fake and deepfake data sequences are set, learning can begin.

2.2.1. Generative learning with GAN

A generative adversarial network (GAN), a neural-network-based unsupervised learning models developed by [10], is implemented to simulate deep fake coin flip sequences. The GAN model consists of generator and discriminator neural networks (NNs) working simultaneously to create realistic synthetic data in an adversarial set up. They have the dual goal to train the generator to produce deepfake sequences that closely mimic real data, fooling the discriminator into falsely classifying them as such, as well as to train the discriminator to detect fake sequences as optimally as feasibly possible.

We design the generator as a feed-forward NN with three dense layers and using Leaky ReLU activations. As an input, a latent vector of 100 independent standard Gaussian (i.e.: $\mathcal{N}(0, 1)$) random variables is taken, which serves as a noise input. Our generator outputs a binary sequence of length 200, each element of which represents a coin flip. At the output layer, to map the results to binary values, we use sigmoid activation. Throughout the network, batch normalization (at varying levels of momentum) and dropout layers (at varying degrees) are used to stabilize the learning process. We train this network using binary cross-entropy loss and the Adam optimizer.

For the weights, we manually initialize them for the generator by drawing the weights from a normal distribution with a standard deviation based on the layer dimensions. The intention of this is to avoid issues with randomness associated with random weight initialization, which allows us to avoid sub-optimal convergence and or local maxima. This initialization ensures the generator starts from a more controlled state, which could potentially improve the quality of the deepfakes it produces.

The discriminator is designed as a more complex classification network. It takes a binary sequence of length 200, which may represent a real sequence, handwritten fake sequence, or any of the three deepfake types from the GAN's generator, MOM's generator, or the Bernoulli generation algorithm.

Like the generator, the discriminator consists of dense layers with Leaky ReLU activations, batch normalization, and dropout to prevent overfitting. Instead of a simple binary output, we design the discriminator to classify between the five categories of sequences using a softmax output layer. We train this network using sparse categorical cross-entropy and the Adam optimizer. This balances stability and training efficiency.

The adversarial tuning of GAN is explained below. The combined initial and adversarial learning of the GAN generator-discriminator is shown in Algorithm C1 (in Appendix C).

2.2.2. MOM learning

As an alternative to GAN, we frame MOM into a GAN-like structure, by replacing both the generator and discriminator neural networks in the traditional GAN framework with MOM components. The MOM generator is then a pairwise Markov chain with a (potentially high-dimensional) hidden component. The rates p and q and initial distribution μ must be learned for each real, fake and deepfake sequence separately, as Figure 1 highlights.

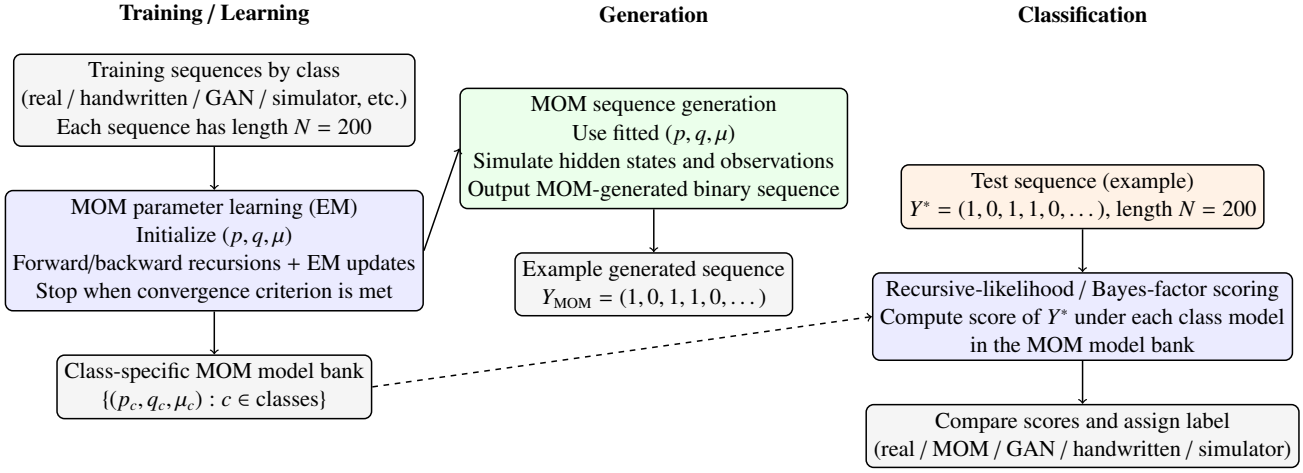

Figure 1. Workflow of the proposed MOM-based framework on a binary-sequence example ($N = 200$). The method consists of three stages: (i) MOM parameter learning via EM (forward/backward recursions and updates of p, q, μ), (ii) sequence generation using fitted MOM parameters, and (iii) classification by recursive-likelihood (Bayes-factor) scoring against class-specific model banks, followed by label assignment.

The algorithm for learning the transition probabilities p and q and the initial distribution μ is given in Algorithm C2 (in Appendix C) and developed in [16]. The specifics of the MOM generator are given below.

The discriminator role in MOM is played by recursive likelihood. The algorithm to compute recursive likelihood is worked out in [16]. The computation and exactly how to use it in the discriminator are stated below.

2.3. GAN's generator

The GAN generator produces deepfake coin flip sequences once fully trained. It takes takes an input vector consisting of 100 independent standard Gaussian random variable samples, and produces an output deemed a deepfake because it is generated by neither a random number generator nor coin flip. The GAN generator is first initially trained as a feed-forward NN and then tuned in the adversarial setup (as described below).

The particulars of the GAN generator neural network are as follows: The input layer is fully-connected with 1024 neurons and a LeakyReLU activation function with a negative slope coefficient of $\alpha = 0.2$. Following this, a dropout layer with a rate of 0.3 is used to prevent overfitting by randomly setting 30% of the input units to 0 during training. The output is then normalized using a batch normalization layer with a momentum of 0.5 to speed up training. Next, the data passes through a second dense layer with 2048 neurons, again with a LeakyReLU activation function ($\alpha = 0.2$), followed by another dropout layer, again with a rate of 0.3, indicating similar regularization. A second batch normalization layer with a momentum of 0.5 is used to normalize the activations. The output is then flattened into a one-dimensional vector by a flatten layer, preparing it for the final dense output layer. The fully-connected output layer has 200 neurons with a sigmoid activation function, ensuring the output values are between 0 and 1, typical for generating binary or normalized data in the range [0, 1]. This neural network is compiled using binary cross-entropy loss and the Adam optimizer with a learning rate of 0.00005.

2.4. MOM's generator

The generator for the MOM solution is a collection of trained pairwise Markov chain models represented by their model parameters $\{p, q, \mu\}$, divided into the groups real (R), fake (F) and deepfake (D). Initially, this collection contains only the models obtained from the initial training data. However, later more models are added from the adversarial setup (as described below). Actual generation occurs by simulating a *randomly-selected real* MOM model with parameters p_R, q_R, μ_R say as shown in Figure 2.

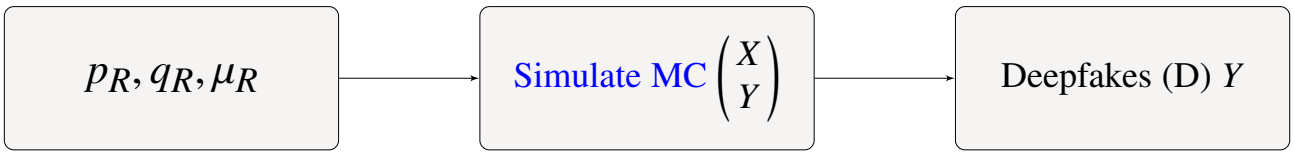

Figure 2. Generation in MOM.

The simulation process consists of simulating the pairwise Markov chain and then discarding the hidden component X to produce the generated observations Y, which are the deepfake coin flips here.

The specifics of the MOM generator are as follows: The observable component Y is just the sequence of coin flips, represented as ones and zeros. The hidden layer X is a Markov chain with s states. We start with $s = 6$. However, during the adversarial process described below, it is allowed to increase. A combined $\begin{pmatrix} X & Y \end{pmatrix}'$ form a Markov chain of dimension $2s$ that has initial distribution

$\begin{pmatrix} X & Y \end{pmatrix}' \sim \mu$ and one step transition probabilities

$$P(X_t = \hat{x}, Y_t = \hat{y} \mid X_{t-1} = x, Y_{t-1} = y) = p_{x\to\hat{x}} q_{y\to\hat{y}}(\hat{x})$$

where p, q, μ have been learned.

This factorization means that the hidden state process X_t follows a Markov chain with transition probabilities $p_{x\to\hat{x}}$, while the observation process Y_t depends on both the current hidden state and the previous observation. In particular, $q_{y\to\hat{y}}(\hat{x})$ depends on the previous observation $Y_{t-1} = y$ and the current hidden state $X_t = \hat{x}$. Thus, unlike a standard HMM, where the observation depends only on the current hidden state, MOM allows the new observation to keep a one-step dependence on the previous observed value [16].

Like many latent-state models, MOM is not uniquely identifiable in a strict sense, since permuting the hidden-state labels can leave the observed distribution unchanged. This issue can become more apparent when we assume the observation process is binary, where different parameter choices may produce very similar observed sequences. For this reason, we treat MOM mainly as a useful latent model for generation and classification, rather than as a uniquely interpretable description of the underlying process. In practice, we reduce these issues by keeping the number of states small and using multiple initializations.

Note that real coin flips could be simulated in this means by making the initial marginal μ_Y for Y fair and then using p, q so that each new Y is equally likely to be 1 or 0 independent of everything. However, there are more efficient ways.

3. Methods of classifying a coin-flip

This section describes different methodologies of classifying coin flips as real or fake. The different classification methods are human, SVM, GAN discriminator classification, BPF method, and the MOM likelihood classification method.

3.1. GAN discriminator

The GAN discriminator, once fully trained, detects (deep)fake coin flip sequences. In particular, the discriminator takes an input vector, consisting of 200 coin flips, and produces an output of one of 5 labels, each corresponding to one of the types of sequences the discriminator is trained on. The discriminator is initially trained as a feed-forward NN and then tuned in the adversarial setup (as described below).

The particulars of the GAN discriminator neural network, optimized to this setting, were determined to be as follows. The input layer is fully-connected with 2048 neurons and a LeakyReLU activation function with a negative slope coefficient of 0.2. Following this, a dropout layer with a rate of 0.6 is used to prevent overfitting by randomly setting 60% of the input units to 0 during training. The output is then normalized using a batch normalization layer with a momentum of 0.5 to stabilize and accelerate training. Next, the data passes through a second dense layer with 1024 neurons, again with a LeakyReLU activation function ($\alpha = 0.2$), followed by another dropout layer, this time with a lower rate of 0.4, indicating less regularization. The fully-connected output layer has 5 neurons with a softmax activation function. The model is compiled with a sparse categorical cross-entropy

loss function and the Adam optimizer with a learning rate of 0.00005 and a β_1 set to 0.5 to smooth out the gradient updates by balancing recent gradient values with past values, which is effective for classification problems and deep learning models due to its adaptive learning rate capabilities.

3.2. Support vector machine classification

In this section, we investigate the support vector machine (SVM) discriminator in classifying real and fake sequences derived from various sources. SVMs have emerged as a powerful tool for such classification tasks due to their effectiveness in high-dimensional data spaces and capability to model both linear and non-linear decision boundaries. In our study, SVM constructs a decision boundary in a high-dimensional space, which is used to classify sequences into "real" and "fake" classes. For SVMs, the "best" decision boundary is typically defined as one that maximizes the margin between the classes. This boundary is determined by support vectors, i.e., data points from each class that are closest to the decision boundary. For linear classification tasks, the boundary can be understood as a hyperplane that maximizes the margin, while in polynomial classification this decision boundary can take more complex, non-linear forms in the input space, adapting to the degree of the kernel used.

In SVM, the best hyperplane is defined as one that maximizes the margin between two classes. This distance is defined as the distance between the hyperplane and the support vectors from either class. The support vectors are the closest data points from either class that have the closest distance to the hyperplane. These support vectors define the elements of the hyperplane and model complex decision boundaries. We evaluated SVM's performance based on accuracy, ROC-AUC, AUC-PR, and F1 scores. This study includes an extensive evaluation based on 100 different random seeds, with results consolidated from multiple iterations to ensure robustness. The 100 random seeds were chosen using `random.randint(1, 100000000)`, which chooses 100 random integers between 1 and 100,000,000.

The data set comprises sequences generated from various methods, including simulator sequences, handwritten sequences, MOM sequences, and GAN sequences, which are all highlighted in the "Methods of Data Generation" section. Additionally, real sequences of length 200 were generated using `random.randint(0,1)` from Python's `random` package, which generates each element in a sequence as a random binary value (0 or 1). We processed each set of sequences into a combined dataset with labels "fake" (0) and "real" (1).

We used polynomial kernels for the classification tasks. The model was trained on the training set subsetted from splitting the full data into an 80-20 split. We evaluated on the testing set using the accuracy, ROC-AUC, AUC-PR, and F1 score metrics. The performance was assessed across varying degrees of polynomial kernels (1, 2, 3, 5, 7, 10, 15) and various random seeds to ensure robustness of the results.

3.3. Filtering classification

The filtering approach as it stands did not require learning. The reason for this is it uses specific real and faker mathematical models. It is true that the faker model has static parameters that could be learnt but these were learnt in an earlier work and regardless, are just numbers that can be learnt one time offline. The generator for the filtering approach is just a random number generator and to include deepfakes the simulator algorithm given above.

We represent the attributes of the faker and the real coin in the mathematical models used in the simulator algorithm and filtering approach. We do this through pairwise covariance and marginal probabilities between each flip and the flips that came before it in time [17]. In filtering terms, the models of the real coin and the fakers with flips have termed signals, and the sequence of coin flips is called observations. We use a filtering technique to generate real-time estimations of the likelihood of various competing signal models based on the observations. We present the problem in a particle filtering framework (see Algorithm C3 (in Appendix C)), describe how we obtained faked data and our observations of its properties, discuss algorithms for simulating flip sequences from marginal probabilities and pairwise covariances, and present empirical results of our implementation of the filtering solution. The fake coin identification problem is presented here within the framework of a filtering algorithm. We need accurate and effective models of both the observations and the possible signals. The signals in this problem are time-inhomogeneous marginal probabilities and covariances, along with a real coin or faker-type indication. We employed the combined branching approach, which is explained in [15]. Classifying, tracking, and predicting the signal based on observations is the aim of filtering. The branching sequential Monte Carlo algorithm was described to reduce particle number fluctuations and thereby improve performance and reliability. We have used three signals $\{X_t^i, t = 1, 2, \dots\}_{i=1}^3$ with the associated weights L_t^i, as mentioned in [17]. σ^N is an approximation of the unnormalized filter in problems like tracking and model selection, which is measured in such a way that $\sigma_t^N = \frac{1}{N} \sum_{j=1}^{N_{t-1}} \hat{L}_t^j \delta_{\hat{X}_t^j}$, where δ denotes the Dirac measure and $\hat{X}_t^j$ denotes the path of the j^{th} particle. In the branching algorithm, when the prior weight $\hat{L}_t^j$ for particle j is outside of a certain interval around the average weight, we do the branching, which helps to preserve the process distribution. In particle filtering, the resampling parameter r and the average particle weight A_t are crucial for maintaining effective particle diversity and avoiding degeneracy, where too few particles carry meaningful weights. The parameter r defines a threshold range around A_t, which is the mean of all particle weights at each time step. This range helps to determine whether particles with significant weight deviations should be resampled. Resampling replicates higher-weight particles while discarding lower-weight ones. When weights are close to A_t, resampling may not be needed, but larger deviations indicate that resampling can help reduce the impact of weight disparity. Adjusting r based on weight variance can further refine the balance between computational efficiency and filter accuracy [4, 9]. In our study, with r set to 4.5, the particle filter uses this specific threshold to determine which particles should be resampled based on how far their weights deviate from the average weight A_t. This means that particles with weights differing significantly (beyond 4.5 times the average weight) are targeted for resampling, which helps maintain a balanced distribution of particle weights and improves the filter's accuracy.

Particles that are branched result in zero or more particles, which are assigned the average weight A_t, and are added at the same location as the parent. In other words, we copy the path with extreme prior weight and give the copies, if there are any, the current average weight. When its prior weight $\hat{L}_t^j$ is not extreme, a particle is not branched and gets to keep its prior weight. The resampling parameter r determines the size of the interval around the average weight A_t, outside of which particles are considered extreme and are branched. The distance between the estimates of value of marginal probabilities and pairwise covariance and the real coin is used as an error metric; a threshold on the error metric is then empirically determined [17]. For a given sequence, if the error metric falls within the threshold, then the sequence is real; otherwise, it is fake.

3.4. MOM's discriminator

The likelihood is a crucial concept in Bayesian statistics, used to quantify the evidence for one model against another. The recursive likelihood from [16] provides a way to update the odds for competing hypotheses based on observed data. The recursive likelihood's ability to incorporate prior information and provide a continuous measure of evidence makes it a versatile and powerful tool for hypothesis testing. One significant advantage of using this is for the interpretability. However, calculating the likelihood can be challenging, especially for complex models with high-dimensional parameter spaces. Advances in computational techniques, such as Markov chain Monte Carlo (MCMC) methods and variational inference, have made it more feasible to compute likelihoods for a wider range of models.

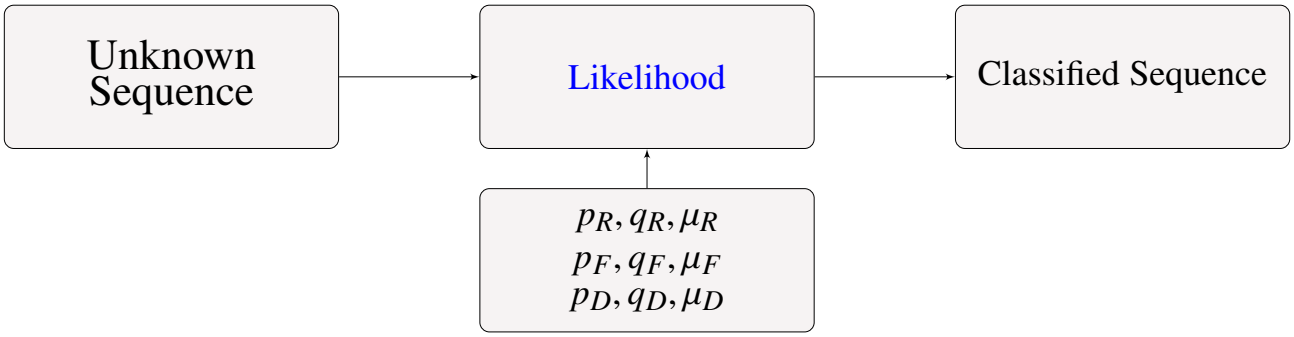

Figure 3. Discriminator in MOM model.

The MOM's discriminator utilizes the recursive likelihood to distinguish the type of sequence being input. For each sequence input into the discriminator, we compute a recursive likelihood between the sequence and every p, q, μ model generated earlier. Each recursive likelihood computed is assigned a label, depending on the type of sequence the p, q, μ model used is from. To classify the sequence, we split the labels and corresponding recursive likelihoods into 2 categories: those that came from a p, q, μ model generated from a sequence of the same type as the input, and the other being the labels that came from the other models. We then take the average of the recursive likelihoods in each group, and classification of the sequence is "correctly identified" or "incorrectly identified", depending on which average is greater.

To integrate the recursive likelihood as a discriminator, we need to follow a process where we calculate the recursive likelihood for each input sequence against pre-trained models and use these factors to classify the sequence as its certain type. Train multiple (p, q, μ) models using the observed sequences. For each input sequence, calculate the recursive likelihood using the pre-trained models. Assume we have the trained models real, GAN, SVM, simulator, and MOM. For an input sequence Y and each (p, q, μ) model from the generation portion, compute the recursive likelihoods of Y with each model. Each recursive likelihood value computed is assigned a label based on the type of sequence the model is known to generate ("Real" for real data sequences, "MOM" for MOM deepfakes, "GAN" for GAN deepfakes, "simulator" for simulator fakes, and "Handfakes" for handwritten fakes). The final classification of the sequence is determined by the maximum between the average of the recursive likelihoods with the correct label, and the average of the recursive likelihoods with the incorrect labels. Should the former be greater, the sequence is classified as "correctly identified", and "incorrectly

identified" otherwise.

By using the recursive likelihood in the MOM discriminator, we leverage the statistical evidence from multiple models to classify sequences. The process involves training multiple models, computing the recursive likelihoods, assigning labels, and using the top recursive likelihood labels to make a final classification decision. This method provides a robust mechanism to distinguish between real and fake sequences based on the collective evidence from several models.

4. Adversarial tuning

As suggested in the name, GAN is an adversarial process which combines the functions of the generator and the discriminator. The GAN takes a latent noise vector as input through the generator, which then generates synthetic sequences. These sequences are passed through the discriminator to classify them as real or fake. Our GAN model is compiled with binary cross-entropy loss and the Adam optimizer. We pre-set the maximum number of epochs to be 500 and the batch size to 64. Within each epoch, the system generates both real and fake samples, trains the discriminator based on these generated samples and the deepfakes, and simultaneously trains both the generator and the discriminator based on their respective loss functions.

For the discriminator, sparse categorical cross-entropy (SCE) loss is used to measure how well the network can classify sequences into one of five categories: real (0), GAN-generated (1), MOM-generated (2), handwritten (3), and simulator (4) sequences. Let N be the number of training samples, x_i the input sequence, $y_i \in \{0, 1, 2, 3, 4\}$ be the true label of the sequence, and $p_\theta(y_i \mid x_i)$ be the discriminator's predicted probability for the correct class, where θ is the current parameter values of the discriminator. The SCE loss is then defined as

$$\mathcal{L}_D = -\frac{1}{N}\sum_{i=1}^{N} \log p_\theta(y_i \mid x_i). \tag{4.1}$$

Given the true labels and the discriminator's predictions, this loss function quantifies the difference between the actual and predicted labels across these five classes. Minimizing $\mathcal{L}_D$ improves the discriminator's ability to correctly classify the type of sequence it receives. By adjusting the network's weights during training, the discriminator learns to more accurately distinguish between real sequences, handwritten sequences, and the various deepfakes.

For the generator, binary cross-entropy (BCE) loss is used. Let z_i be the random latent noise input, $G(z_i)$ be the generator output (i.e. GAN-generated), and $D(G(z_i))$ be the discriminator's probability that $G(z_i)$ is real. The BCE is

$$\mathcal{L}_G = -\frac{1}{N}\sum_{i=1}^{N} \left[y_i \log(D(G(z_i))) + (1 - y_i) \log(1 - D(G(z_i))) \right]. \tag{4.2}$$

Here, the generator's goal is to produce realistic deepfake sequences to fool the discriminator into outputting a high probability of them being real, close to 1. Therefore, the BCE loss of the generator is calculated with "opposite" labels, where it aims to minimize the loss between its fake output (which it wants the discriminator to classify as real) and the real label.

Minimizing each network's respective loss functions essentially leads to a more effective adversarial relationship between the generator and the discriminator. This drives the GAN to improve both the

generator's ability to synthesize realistic data and the discriminator's ability to correctly classify the data as real or fake. As per [11], GAN training is usually expressed as a min-max game:

$$\min_G \max_D V(D,G) = \mathbb{E}_{x\sim p_{\text{data}}}[\log(D(x))] + \mathbb{E}_{z\sim p_z}[\log(1 - D(G(z)))].$$

The model is compiled with sparse categorical cross-entropy (SCE) loss for the discriminator and binary cross-entropy (BCE) loss for the generator, both of which are optimized using the Adam optimizer. The use of SCE allows the discriminator to classify sequences into one of five categories, while BCE ensures that the generator's goal is to produce deepfake data that is realistic enough to fool the discriminator into classifying the fake data as real. During training, the discriminator should ideally output a high probability for real data and a low probability for the fake sequences generated by the GAN.

5. Comparative results

In this section, Table 1 presents how the real, handwritten, and sequences generated from simulator, GAN, and MOM are classified using the GAN discriminator, support vector machines (SVM), particle filtering, and the MOM recursive likelihood approach. In other words, we want to study whether MOM performs better in classifying different types of sequences than the other methods. All methods perform multiclass classification over the five sequence types: real, handwritten, simulator, GAN, and MOM. For each method, we perform studies for the handwritten fake sequences, and the other generated sequences. There are 137 fake sequences of each type, each of length 200. The number of real sequences we generate corresponds to however many fake sequences we consider (either 137 or 548).

Across different polynomial degrees, the SVM model performs best with lower-degree kernels (1, 2, and 3) for generating sequences like GAN and handwritten sequences, with some variation in performance on more complex types like simulator and MOM sequences. The overall accuracy improves up to degree 2 but fluctuates afterwards, with degree 5 showing some promise. However, higher degrees (7 and 10) give worsening results, suggesting overfitting and increased complexity that the model cannot handle effectively. Standard deviations are consistently low, indicating that the model performs consistently across runs, but it struggles to generalize across all sequence types. The results suggest that while SVM can be tuned to achieve higher accuracy on specific types of sequences, its overall performance across different types of fakes is relatively stable but not particularly high.

For GAN's training, we split the types of sequences into training and testing sets using an 80:20 split. We train on 2000 epochs using a batch size of 64. For testing, we feed each sequence in the testing set into the trained discriminator. In Table 1, the GAN shows mixed results. It performs poorly on the Bernoulli generation algorithm (6.18%) and MOM (7.36%) sequences, indicating difficulty in accurately detecting these types. However, it achieves moderate success on GAN sequences (54.96%) and performs reasonably well on handwritten sequences (74.64%) and real sequences (74.33%). Overall, the GAN model achieves 54.88% accuracy, but the relatively high standard deviations (ranging from 2.87 to 9.41) suggest inconsistent performance, making the model less reliable across different sequence types . This indicates that while the GAN discriminator is moderately effective at distinguishing real sequences, its ability to detect fake sequences, particularly handwritten ones, is less robust, showing significant variability in its performance.

Table 1. Classification performance by sequence type in the 5-class setting. The sequence types are: *real*, computer-generated fair coin-flip sequences treated as authentic; *Handwritten Seq.*, human-produced fake coin-flip sequences; *Simulator*, deepfake sequences produced by the Bernoulli-generation simulator; *GAN*, deepfake sequences generated by the GAN generator; and *MOM*, deepfake sequences generated by the MOM. For each method, the entries in the five sequence-type columns report the held-out classification accuracy for sequences whose true label is the indicated class. The *Overall* column reports pooled accuracy on the combined held-out evaluation set across all five sequence types, not a macro-average.

	Simulator (%)	GAN (%)	Handwritten Seq. (%)	MOM (%)	Real (%)	Overall (%)
SVM						
Degree 1						
Accuracy	53.31	73.89	60.98	**50.1**	51.09	47.42
Standard deviation	0.05	0.02	0.05	0.03	0.08	0.06
Degree 2						
Accuracy	53.32	71.33	69.33	**50.4**	58.81	52.7
Standard deviation	0.05	0.02	0.08	0.03	0.05	0.01
Degree 3						
Accuracy	47.43	75	70.06	**51.5**	50.96	49.34
Standard deviation	0.06	0.04	0.06	0.01	0.01	0.01
Degree 5						
Accuracy	51.82	74.27	71.19	**48.6**	49.2	50.88
Standard deviation	0.03	0.07	0.05	0.04	0.02	0.01
Degree 7						
Accuracy	52.59	74.63	52.59	**48.9**	49.67	50.83
Standard deviation	0.04	0.04	0.03	0.03	0.02	0.01
Degree 10						
Accuracy	49.26	53.29	51.09	**47.1**	49.7	51.82
Standard deviation	0.01	0.02	0.01	0.05	0.02	0.01
GAN						
Accuracy	6.18	54.96	74.64	**7.36**	74.33	54.88
Standard deviation	4.89	9.41	5.91	4.42	5.11	2.87
Particle filtering						
Accuracy	86.86	80.67	**90.88**	**73.67**	87.96	86.32
Standard deviation	0.025	0.3	0.012	0.024	0.018	0.010
MOM						
Accuracy	**99.82**	**97.25**	89.29	**76.59**	**98.76**	**92.27**
Standard deviation	0.78	3.89	2.43e-13	13.53	2.05	2.93

The particle filtering applied to different coin flip sequences, specifically handwritten and all generated (MOM, GAN, Bernoulli generation algorithm, and handwritten) sequences. The average accuracy is used to evaluate the performance of the particle filtering algorithm. Particle filtering demonstrates strong and consistent performance across all sequence types. With accuracies ranging from 80.67% (handwritten sequences) to 90.88% (GAN sequences), it shows proficiency in handling a variety of sequence types. The overall accuracy is 86.32%, making it one of the top-performing models. Importantly, the model has very low standard deviations (between 0.08 and 0.09 for each type), indicating highly stable and reliable performance. Overall, these results highlight the efficacy of the particle filtering algorithm in accurately distinguishing between different types of coin flip sequences, maintaining high accuracy and low variability across both handwritten and other fake sequences. This consistency in performance underscores the algorithm's potential reliability and effectiveness in practical applications involving sequence analysis and classification.

For MOM, similar to the GAN, we split the sequences into training and testing sets using an 80:20 split. We first consider a hidden layer consisting of s states. To initiate our model, we consider a "canonical model" and consider a case where the hidden layer has 2 states and a q matrix that mimics a "perfect" coin flip (where all entries are 0.25). Consider a randomly generated p matrix and a μ matrix with 0.25 as every entry. Using these matrices, we generate the canonical model and use this to initialize the set of real sequences. The rest of the real sequences are randomly generated using Python's `random` package. Using the EM algorithm outlined in [16], the p, q, μ transition matrices of the real sequences generated by Python's `random` module are maximized starting from 6 hidden states. Then, a subset of the maximized p, q, μ matrices of the real sequences are used to simulate more coin flip sequences, which we refer to as "generated deepfakes" data. Post-generation, we increase the hidden states to 7 and re-generate new p, q, μ models of the real sequences, while generating the p, μ matrices of the generated and fake sequences following the same procedure as outlined above. The Algorithm C2 describes how to produce real, generated, and fake sequences. Note that fake sequences are created by hand to mimic coin flips as realistic as possible. Here, we present an algorithm to generate and analyze sequences of coin flips as realistically as possible.

For the training portion of the MOM, we initially consider a hidden layer consisting of six states. After 100 trials running against sequences, the MOM exhibits the highest overall accuracy, with an impressive 92.27%. It performs exceptionally well on the Bernoulli generation algorithm (99.82%), GAN (97.25%), and real (98.76%) sequences, showing that it is highly effective in detecting most types of sequences. However, its accuracy on MOM sequences is lower (76.59%), suggesting some difficulty in detecting sequences of its own type, which demonstrates the strength of the MOM generation. While the model generally performs well, the higher standard deviation for MOM sequences (13.53) indicates variability in its performance for this category. These results indicate that MOM is not only highly effective at distinguishing between real and fake sequences, but also particularly strong at deepfake detection, especially compared to GAN. Table 1 provides the results of the performances of each classification method against each type of sequence, as well as the overall accuracy of each method.

Complete implementation details for GAN, MOM, and BPF (including hyperparameters, initialization, stopping criteria, and runtime settings used for Table 1) are provided in Appendix A.

5.1. Experimental protocol

Each sequence has length $T = 200$. We consider five sequence types: real, handwritten fake, and fake sequences generated by simulator, GAN, and MOM. For each fake type, we construct a balanced binary classification dataset by matching the number of real sequences 1:1 to the number of fake sequences (137 real and 137 fake sequences). For pooled experiments across all four fake types, we use 548 real sequences to match the 548 fake sequences.

Unless otherwise stated, all classifiers (SVM, GAN discriminator, particle filtering, and MOM discriminator) use the same stratified 80:20 train/test split at the sequence level. For each repeated run, the train/test split is generated independently, and all methods are evaluated on the same split for fairness. Hyperparameters are selected using training data only (via a validation split within the training set), and the test set is used only for final evaluation.

We report mean accuracy and standard deviation over 100 repeated runs with different random seeds.

5.2. Human baseline mini-study

To provide a baseline for comparison with machine learning methods, we conducted a small study observing human performance on the coin flip discrimination task. A total of 83 respondents were presented with 30 binary sequences: 10 *real*, 10 *handwritten*, and 10 *MOM generated deepfakes*. For each sequence, respondents were asked to classify it as real, handwritten, or deepfake.

Figure 4 shows the distribution of total quiz scores. The mean score was 11/30 (median = 11), only slightly above chance performance. The majority of participants only scored between 10 and 17, with no score being higher than 17, and the lowest score being 4.

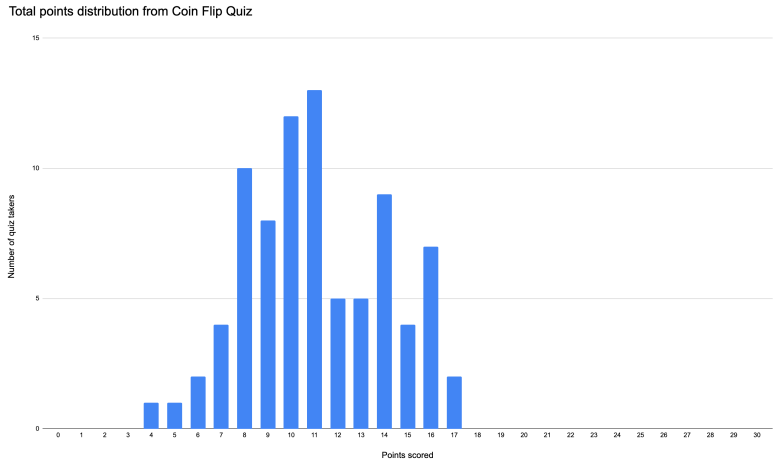

Figure 4. Distribution of the results of the coin flip quiz.

The aggregate classification results are summarized in the confusion matrix Table 2. Each row corresponds to the true sequence type, and each column corresponds to the human classification. The rows each sum to 830, which reflects the 10 sequences of each type evaluated by the 83 respondents.

Table 2. Confusion matrix of the aggregate quiz results.

True/predicted	Predicted real	Predicted handwritten	Predicted deepfake
True real	320	260	250
True handwritten	236	333	261
True deepfake	310	252	268

Notice the emergence of several patterns. Respondents were most successful at identifying human-written fakes (around 43.3% accuracy), which suggests that such sequences contain detectable artifacts to the human eye. In contrast, performance was much poorer on distinguishing real from deepfake sequences: Only around 40.2% of real sequences were correctly identified, with nearly as many (around 31.4%) misclassified as deepfakes. Similarly, only around 35.2% of deepfakes were correctly labeled, with many misclassified as real (38.1%).

In the end, participants did perform above chance, but far from reliably so. Their relative success at spotting handwritten fakes was contrasted with their near chance accuracy in distinguishing real sequences from deepfakes. Here, we draw the conclusion that the human eye is poorly suited for detecting algorithmically generated randomness, which emphasizes the value of the existence of probabilistic and machine learning methods for deepfake detection.

6. Conclusions and future work

Based on our comparative analysis, several key insights emerge regarding the creation and detection of deepfakes using different classification methods. The GAN discriminator, while moderately effective, demonstrates significant variability in its ability to detect fake sequences, particularly handwritten ones. Its overall accuracy drops considerably when tested against all types of fake sequences, indicating limitations in generalizability and robustness. This suggests that while GANs can generate realistic sequences, their discriminators require further refinement to reliably distinguish between real and fake data. The SVM's performance improves with the degree of the polynomial to a point, achieving the highest accuracy with a degree of 5 for handwritten sequences. However, its performance declines sharply with higher degrees and remains relatively stable but not particularly high when tested against all types of fake sequences. This indicates that while SVMs can be fine-tuned for specific types of deepfake detection, their overall efficacy across diverse fake sequences is limited. The particle filtering method shows impressive accuracy and consistency in classifying both handwritten and all types of fake sequences. Its high accuracy and low variability underscore its potential reliability and effectiveness in practical applications involving sequence analysis and classification. This method demonstrates robustness in detecting deepfakes generated by various methods, making it a valuable tool in the fight against deepfake fraud. MOM outperforms the other methods significantly, achieving near-perfect accuracy for both real and handwritten sequences and maintaining high performance against all types of fake sequences. Its minimal performance variability and consistently high accuracy highlight its robustness and efficacy in distinguishing between real and fake sequences.

In addition to accuracy, the methods differ substantially in computational cost. The asymptotic complexity analysis in Appendix B shows that the BPF scales linearly in sequence length and particle count, while the GAN has a higher up-front training cost but relatively cheap classification cost once the discriminator is trained. By contrast, MOM is computationally heavier than both because it

combines repeated EM fitting, multiple random initializations, and model-bank scoring at test time. In our implementation, this additional computational cost was accompanied by stronger classification performance, indicating a practical trade-off between efficiency and accuracy in the controlled setting considered here.

In summary, while GAN offers some utility in deepfake detection, its performance is less reliable and variable. Particle filtering shows robust performance, and MOM stands out as the most effective method, demonstrating high accuracy and consistency. Future work should focus on enhancing the robustness and generalizability of GANs and SVMs, exploring the integration of particle filtering and MOM techniques, and developing new methods to improve deepfake detection accuracy further. Additionally, research should consider the evolving complexity of deepfake generation techniques to ensure detection methods remain effective against increasingly sophisticated fakes. Further exploration into hybrid models combining strengths from multiple techniques could also provide more comprehensive solutions for deepfake detection.

Author contributions

Michael Kouritzin developed models and provided general supervision over the project. Ian Zhang was tasked with the GAN and MOM sections, running the comparative results and the human benchmark mini-study. Jyoti Bhadana handled the adversarial tuning and branching particle filtering, and Seoyeon Park was tasked with the SVM part. All authors have read and approved the final version of the manuscript for publication.

Use of Generative-AI tools declaration

The authors declare that they have not used Artificial Intelligence (AI) tools in the creation of this article.

Conflict of interest

Prof. Michael A. Kouritzin is the Guest Editor of special issue "The Mathematics of Artificial Intelligence" for AIMS Mathematics. Prof. Michael A. Kouritzin was not involved in the editorial review and the decision to publish this article.

All authors declare no conflicts of interest in this paper.

Appendices

A. Implementation details and reproducibility

A.1. GAN implementation details

The GAN discriminator is trained as a 5-class classifier with labels

$$0 = \text{real}, \quad 1 = \text{GAN}, \quad 2 = \text{MOM}, \quad 3 = \text{handwritten}, \quad 4 = \text{simulator}.$$

Let x_i denote an input sequence and $y_i \in \{0, 1, 2, 3, 4\}$ its class label. The discriminator parameters θ_D are trained using sparse categorical cross-entropy:

$$\mathcal{L}_D(\theta_D) = -\frac{1}{N} \sum_{i=1}^{N} \log p_{\theta_D}(y_i \mid x_i).$$

In each discriminator update, the loss is computed separately for real, GAN-generated, handwritten, simulator, and MOM-generated mini-batches, and the five losses are averaged.

The generator is trained through the combined GAN model using sparse categorical cross-entropy with target label 0 (the "real" class) (i.e., it is optimized to generate sequences that the discriminator classifies as real).

The generator takes a latent vector of dimension 100 and outputs a binary sequence of length 200 (via a sigmoid output layer followed by thresholding at 0.5 for generation). The generator uses two hidden dense layers (1024 and 2048 units) with LeakyReLU activations, dropout (rate 0.3), and batch normalization (momentum 0.5). The discriminator uses dense layers of sizes 2048 and 1024 with LeakyReLU activations, dropout (rate 0.3), and batch normalization (momentum 0.5), followed by a 5-class softmax output.

For the final experiments reported in Table 1, we used:

- Adam optimizer for both generator and discriminator,
- learning rate 10^{-5},
- $\beta_1 = 0.5$,
- batch size 64,
- 500 epochs,
- one discriminator update and two generator updates per epoch.

No learning-rate schedule was used.

To improve training stability, we used dropout and batch normalization in both networks, and added Gaussian noise with standard deviation 0.1 to real and generated sequences during discriminator training. We did not use additional anti-collapse penalties (e.g., gradient penalty or feature matching).

The GAN generator outputs a length 200 vector with sigmoid activations. To obtain a binary sequence, each coordinate is thresholded at 0.5: values greater than 0.5 are mapped to 1, and values less than or equal to 0.5 are mapped to 0. No temperature scaling, Gumbel-softmax, or other stochastic sampling rule is used at generation time. The latent input is sampled i.i.d. from a 100-dimensional standard Gaussian distribution.

During discriminator training, additive Gaussian noise with standard deviation 0.1 is applied to both real and GAN-generated sequences before the discriminator update. This same noise setting is used in all runs. We did not use a separate temperature-based or stochastic label-smoothing schedule.

A.2. MOM implementation details

The MOM experiments were run over 100 random seeds. For each run, the hidden-state dimension was initialized at $s = 6$ and the observation was set to be binary. The state-size schedule was fixed (not adaptively selected): we first fit models with $s = 6$ to generate MOM-based fake sequences, then increased the hidden-state dimension once to $s = 7$ and refit class-specific models used for classification.

In the MOM implementation, real sequences were modeled with length $N_{\text{real}} = 400$, while fake sequences (MOM-generated, handwritten, tricky, and GAN-generated) were modeled with length $N_{\text{fake}} = 200$. We used an 80:20 train/test split for each class.

For each fitted model:

- P was initialized as a random row-stochastic matrix by sampling each row from i.i.d. Uniform(0, 1) values and normalizing;
- Q was initialized from empirical transition counts of the observed sequence, followed by random perturbation of nonzero entries and row normalization;
- μ was randomly initialized, with entries incompatible with the first observation forced to zero, followed by normalization to sum to one.

Parameter updates were performed using forward-backward recursions and iterative updates of (P, Q, μ). For each fit, we used a maximum of 1000 iterations. The stopping criterion was based on parameter changes (not log-likelihood): we stopped when

$$\sum_{i,j} |P_{ij}^{(m)} - P_{ij}^{(m-1)}| < 5 \times 10^{-3} \quad \text{and} \quad \sum_{i,y} |\mu_{iy}^{(m)} - \mu_{iy}^{(m-1)}| < 5 \times 10^{-3}.$$

For each training sequence and each class, we generated 10 random initializations (denoted `model1`-`model10`) and fit the corresponding MOM parameters. For each fitted model, we computed a recursive likelihood. In the implementation used for the reported experiments, the score was recorded for all 10 fits and the final fitted model in the loop was retained for later usage.

For each test sequence, we evaluated Bayes-factor scores against all fitted model banks (real, MOM-generated, handwritten, simulator, and GAN). Class-specific decisions were made by comparing the mean recursive likelihood over models from the target class against the mean recursive likelihood over models from the remaining classes.

A.3. Branching particle filter (BPF) implementation details

For the branching particle filter, we initialized the filter with 200 particles and averaged the reported results over 100 independent runs to reduce Monte Carlo variability.

Branching was controlled using a threshold parameter r in the range $3 \leq r \leq 3.5$ (as used in the interval test in Algorithm C3). In our experiments, this range provided stable performance.

We also tested larger particle counts and observed little improvement in accuracy relative to 200 particles, while computational cost increased noticeably. For this reason, we used 200 particles in the final reported experiments.

B. Runtime and computational cost

For the GAN (shown in Algorithm C1), let

- E be the number of epochs;
- M be the number of mini-batches per epoch;
- B be the batch size;
- N be the sequence length ($N = 200$);
- d_z be the latent dimension ($d_z = 100$);

- C be the number of classes in the discriminator ($C = 5$).

Let W_G and W_D denote the per-sample dense-layer operation counts for generator and discriminator. In our architecture,

$$W_G = O(d_z \cdot 1024 + 1024 \cdot 2048 + 2048 \cdot N),$$
$$W_D = O(N \cdot 2048 + 2048 \cdot 1024 + 1024 \cdot C).$$

Each discriminator update processes five class-specific batches (real, GAN, MOM, handwritten, simulator), giving $O(BW_D)$ work per mini-batch up to a constant factor. A generator update through the combined GAN model costs $O(B(W_G + W_D))$. Therefore, the total GAN training complexity is

$$O(E(MBW_D + B(W_G + W_D))).$$

For the MOM (shown in Algorithm C2), let

- N be the sequence length;
- s be the hidden-state dimension;
- K be the number of EM iterations (until stopping, capped by the stopping limit);
- R be the number of random initializations per sequence (in our implementation, $R = 10$);
- J_1 be the number of sequences fit in the initial state-s stage;
- J_2 be the number of sequences fit after raising the state dimension to $s + 1$;
- L be the number of stored MOM models used for recursive-likelihood (Bayes-factor) scoring at test time.

In one EM iteration, the dominant cost comes from the forward and backward recursions and the parameter updates for p, q, μ. Each of these scales as $O(Ns^2)$. Therefore, one EM iteration costs

$$O(Ns^2).$$

Over K EM iterations, fitting one MOM model costs

$$O(KNs^2).$$

Since each sequence is fit from R random initializations, the cost per sequence becomes

$$O(RKNs^2).$$

Algorithm C2 has two fitting stages: one at state dimension s, and one after increasing the state dimension to $s + 1$. Therefore, the total MOM fitting cost is

$$O(J_1RKNs^2 + J_2RKN(s + 1)^2).$$

At test time, recursive likelihood scoring of one sequence against L stored MOM models costs

$$O(LN(s + 1)^2),$$

since classification is performed after the state dimension is increased.

For the BPF (shown in Algorithm C3), let

- N be the sequence length (number of time steps);
- P_t be the number of particles at time step t;
- P be a representative particle count (e.g., average or typical particle count over time);
- R_{PF} be the number of independent BPF runs (in our implementation, $R_{\mathrm{PF}} = 100$).

At each time step t, the algorithm:

- evolves particles independently;
- computes weighted summaries and the average weight;
- checks which particles branch;
- generates offspring for branched particles.

Thus, the work at time step t is proportional to

$$O(P_{t-1} + P_t).$$

Summing over all N time steps, one BPF run has complexity

$$O\left(\sum_{t=1}^{N}(P_{t-1} + P_t)\right).$$

If the particle count is kept approximately stable (as intended in the branching scheme), i.e. $P_t \approx P$, this simplifies to

$$O(NP).$$

Averaging over R_{PF} independent runs gives total BPF cost

$$O(R_{\mathrm{PF}}NP).$$

In our implementation, we used 200 particles and averaged over 100 runs, so this is $O(100NP)$ up to constant factors.

The three methods have different computational profiles, so we compare them separately in terms of fitting/training and test-time scoring (see Table B). The GAN has a higher up-front optimization cost (epochs, mini-batches, and backpropagation through the generator/discriminator), but test-time classification is cheap once the discriminator is trained (one forward pass, $O(W_D)$ per sequence). In contrast, MOM has heavier fitting cost due to repeated EM iterations and multiple random initializations, and its test-time scoring cost is also larger because each test sequence is compared against a bank of fitted models. The BPF scales linearly in sequence length and particle count, and is the lightest asymptotically per run among the three methods.

Table B. Asymptotic complexity comparison.

Method	Training / fitting cost	Test-time scoring cost
GAN	$O(E(MBW_D + B(W_G + W_D)))$	$O(W_D)$ per sequence
MOM	$O(J_1RKNs^2 + J_2RKN(s+1)^2)$	$O(LN(s+1)^2)$ per sequence
BPF	$O(R_{\mathrm{PF}}NP)$	$O(NP)$ per run/sequence

This asymptotic comparison is consistent with our implementation-level observation that BPF is computationally cheaper than MOM in our setting, while MOM incurs additional computational cost due to repeated EM fitting and model-bank scoring. In this sense, MOM trades computational time for improved classification performance in our controlled experiments.

C. Algorithms

Algorithm C1: GAN training process for coin-flip sequences.

Data: Real coin flip sequences
Input: Initialized generator G and discriminator D with random weights
Function `TrainGAN`($G, D, epochs, batch\_size$)
  for $epoch = 1$ **to** $epochs$ **do**
    // Train generator: prepare latent input
    $z \leftarrow$ generate_latent_points($batch\_size$)
    $y_{\text{gan}} \leftarrow \mathbf{1}$ // inverted labels for fake samples
    // Train discriminator
    foreach *batch* **do**
      // Get real coin-flip sequences
      $(X_{\text{real}}, y_{\text{real}}) \leftarrow$ generate_real_samples($batch\_size$)
      // Generate GAN deepfake sequences
      $z \leftarrow$ generate_latent_points($batch\_size$)
      $X_{\text{GAN}} \leftarrow G(z)$
      $y_{\text{GAN}} \leftarrow \mathbf{1}$
      // Get MOM deepfake sequences and labels
      $(X_{\text{MOM}}, y_{\text{MOM}}) \leftarrow$ generate_MOM_samples($batch\_size$)
      // Get handwritten fake sequences and labels
      $(X_{\text{handfake}}, y_{\text{handfake}}) \leftarrow$ generate_handwritten_samples($batch\_size$)
      // Get simulator deepfake sequences and labels
      $(X_{\text{simulator}}, y_{\text{simulator}}) \leftarrow$ generate_simulator_samples($batch\_size$)
      // Train discriminator on real and fake samples
      $d_{\text{loss,real}} \leftarrow D$.train_on_batch($X_{\text{real}}, y_{\text{real}}$)
      $d_{\text{loss,GAN}} \leftarrow D$.train_on_batch($X_{\text{GAN}}, y_{\text{GAN}}$)
      $d_{\text{loss,MOM}} \leftarrow D$.train_on_batch($X_{\text{MOM}}, y_{\text{MOM}}$)
      $d_{\text{loss,handfake}} \leftarrow D$.train_on_batch($X_{\text{handfake}}, y_{\text{handfake}}$)
      $d_{\text{loss,simulator}} \leftarrow D$.train_on_batch($X_{\text{simulator}}, y_{\text{simulator}}$)
    // Update generator through discriminator error
    $g_{\text{loss}} \leftarrow GAN$.train_on_batch($z, y_{\text{gan}}$)
    // Print progress every 100 epochs
    if *epoch % 100 == 0* **then**
      // Display current training status
      print "Epoch", epoch
      print "Discriminator Loss:",
        $(d_{\text{loss,real}} + d_{\text{loss,GAN}} + d_{\text{loss,MOM}} + d_{\text{loss,handfake}} + d_{\text{loss,simulator}})/5$
      print "Generator Loss:", g_{loss}

Algorithm C2: MOM training process.

Data: Observations of real and fake coin sequences, Y

Input: Initial transition probabilities $p_{x\to x'}$, $q_{y\to y'}(x)$, and initial distribution $\mu(x,y)$

Function *TrainMOM*(p, q, μ, N, Y)

foreach *real training sequence i* **do**

`// Forward propagation`

$$\pi_0(x,y) \leftarrow \mu(x,y)$$

$$\pi_1(x) \leftarrow \frac{\sum_{x_0\in E}\sum_{y_0\in O}\mu(x_0,y_0)\,p_{x_0\to x}\,q_{y_0\to Y_1}}{\sum_{x}\left(\sum_{x_0\in E}\sum_{y_0\in O}\mu(x_0,y_0)\,p_{x_0\to x}\,q_{y_0\to Y_1}\right)}$$

for $n = 2, 3, \ldots, N$ **do**

$$\pi_n(x) \leftarrow \frac{q_{Y_{n-1}\to Y_n}(x)\sum_{x_{n-1}}\pi_{n-1}(x_{n-1})p_{x_{n-1}\to x}}{\sum_{x} q_{Y_{n-1}\to Y_n}(x)\sum_{x_{n-1}}\pi_{n-1}(x_{n-1})p_{x_{n-1}\to x}}$$

`// Backward propagation`

$$\chi_{N-1}(x) \leftarrow q_{Y_{N-1}\to Y_N}(x)$$

$$\chi_0(x,y) \leftarrow \frac{q_{y\to Y_1}(x)}{\sum_{x}\left(\sum_{x_0\in E}\sum_{y_0\in O}\mu(x_0,y_0)\,p_{x_0\to x}\,q_{y_0\to Y_1}\right)}$$

for $n = N-1, N-3, \ldots, 1$ **do**

$$\chi_n(x) \leftarrow \frac{q_{Y_n\to Y_{n+1}}(x)}{\sum_{x} q_{Y_n\to Y_{n+1}}(x)\sum_{x_n}\pi_n(x_n)p_{x_n\to x}}$$

`// EM updates in state` s

$$p_{x\to x'} \leftarrow \frac{p_{x\to x'}\left[\sum_{y}\pi_0(x,y)\chi_0(x',y) + \sum_{n=1}^{N-1}\pi_n(x)\chi_n(x')\right]}{\sum_{x_1} p_{x\to x_1}\left[\sum_{y}\pi_0(x,y)\chi_0(x_1,y) + \sum_{n=1}^{N-1}\pi_n(x)\chi_n(x_1)\right]}$$

$$q_{y\to y'}(x) \leftarrow \frac{\sum_{\xi} p_{\xi\to x}\left[\mathbf{1}_{\{Y_1=y'\}}\chi_0(x,y)\pi_0(\xi,y) + \sum_{n=1}^{N-1}\mathbf{1}_{\{Y_n=y,\,Y_{n+1}=y'\}}\chi_n(x)\pi_n(\xi)\right]}{\sum_{\xi} p_{\xi\to x}\left[\chi_0(x,y)\pi_0(\xi,y) + \sum_{n=1}^{N-1}\mathbf{1}_{\{Y_n=y\}}\chi_n(x)\pi_n(\xi)\right]}$$

$$\mu(x,y) \leftarrow \frac{\mu(x,y)\sum_{x_1}\chi_0(x_1,y)p_{x\to x_1}}{\sum_{\xi}\sum_{\theta}\mu(\xi,\theta)\sum_{x_1}\chi_0(x_1,\theta)p_{\xi\to x_1}}$$

`// Simulate a coin sequence` j `from the fitted real-sequence model`

Simulate coin sequence j using $p_{i,\text{real}}$, $q_{i,\text{real}}$, and $\mu_{i,\text{real}}$

foreach *model* $(p_{i,\text{real}}, q_{i,\text{real}}, \mu_{i,\text{real}})$ **do**

Compute recursive likelihood between the input sequence and the model

`// Classify by largest recursive likelihood`

Compare recursive likelihoods and assign the label with highest score

`// Raise state dimension and refit`

Raise state dimension from s to $s+1$

foreach *real training sequence i* **do**

Re-compute final estimates of $p_{i,\text{real}}$, $q_{i,\text{real}}$, and $\mu_{i,\text{real}}$ in state $s+1$ via EM

foreach *other sequences (MOM, GAN, handwritten, simulator)* **do**

Compute recursive likelihoods and estimate p, q, and μ in state $s+1$ via EM

Algorithm C3: Branching algorithm.

Function *BranchingAlgorithm*$(X_0, L_0, N, T, \{r_t\}_{t=1}^{T})$

for $t = 1$ **to** T **do**

// Evolve particles independently

$(X_{t-1}^{j}, L_{t-1}^{j}) \leftarrow (\hat{X}_t^j, \hat{L}_t^j)$

// Estimate σ_t

$S_t^N \leftarrow \dfrac{1}{N} \sum_{j=1}^{N_{t-1}} \hat{L}_t^j \, \delta_{\hat{X}_t^j}$

// Calculate average weight

$A_t \leftarrow S_t^N(1)$

// Check which particles branch

$m \leftarrow 0$

for $j = 1$ **to** N_{t-1} **do**

if $\hat{L}_t^j \in \left(\dfrac{1}{r_t} A_t, \; r_t A_t\right)$ **then**

// Move non-branched particles to final vector

$(X_t^{j-m}, L_t^{j-m}) \leftarrow (\hat{X}_t^j, \hat{L}_t^j)$

else

$m \leftarrow m + 1$

$(X_t^m, L_t^m) \leftarrow (\hat{X}_t^j, \hat{L}_t^j)$

// Branching part of the algorithm

$N_t \leftarrow m$

Simulate $\{V_t^j\}_{j=m+1}^{N_{t-1}}$ independently with

$$V_t^j \sim \text{Uniform}\left[\frac{j-m-1}{N_{t-1}-m}, \frac{j-m}{N_{t-1}-m}\right]$$

Let p be a random permutation of $\{m+1, m+2, \ldots, N_{t-1}\}$

for $j = m + 1$ **to** N_{t-1} **do**

$U_t^j \leftarrow V_t^{p(j)}$

for $j = m + 1$ **to** N_{t-1} **do**

$$N_t^j \leftarrow \left\lfloor \frac{\hat{L}_t^{j-m}}{A_t} \right\rfloor + \mathbf{1}\left\{ U_t^j \leq \left(\frac{\hat{L}_t^{j-m}}{A_t} - \left\lfloor \frac{\hat{L}_t^{j-m}}{A_t} \right\rfloor \right) \right\}$$

for $k = 1$ **to** N_t^j **do**

$(X_t^{N_t+k}, L_t^{N_t+k}) \leftarrow (\hat{X}_t^{j-m}, A_t)$

$N_t \leftarrow N_t + N_t^j$

References

1. D. J. Ballantyne, H. Y. Chan, M. A. Kouritzin, Novel branching particle method for tracking, *Signal and Data Processing of Small Targets 2000*, **4048** (2000), 277–287. https://doi.org/10.1117/12.391984
2. T. Baltrusaitis, C. Ahuja, L. P. Morency, Multimodal machine learning: a survey and taxonomy, *IEEE Trans. Pattern Anal. Mach. Intell.*, **41** (2019), 423–443. https://doi.org/10.1109/tpami.2018.2798607
3. C. Batanero, E. Sanchez, What is the nature of high school students' conceptions and misconceptions about probability? In: G. A. Jones, *Exploring probability in school: challenges for teaching and learning*, Mathematics Education Library, Springer US, Boston, MA, 2005, 241–266. https://doi.org/10.1007/0-387-24530-8_11
4. O. Cappe, S. J. Godsill, E. Moulines, An overview of existing methods and recent advances in sequential Monte Carlo, *Proceedings of the IEEE*, **95** (2007), 899–924. https://doi.org/10.1109/JPROC.2007.893250
5. M. Costa, L. De Angelis, *Model selection in hidden markov models: a simulation study*, Quaderni di Dipartimento 7, Department of Statistics, University of Bologna, 2010.
6. D. Crisan, T. Lyons, Nonlinear filtering and measure-valued processes, *Probab. Theory Relat. Fields*, **109** (1997), 217–244. https://doi.org/10.1007/s004400050131
7. P. Del Moral, L. Miclo. Branching and interacting particle systems approximations of feynman-kac formulae with applications to non-linear filtering, In: J. Azéma, M. Ledoux, M. Émery, M. Yor, *Séminaire de probabilités XXXIV*, Berlin, Heidelberg: Springer, **1729** (2000), 1–145. https://doi.org/10.1007/bfb0103798
8. R. Douc, O. Cappe, Comparison of resampling schemes for particle filtering, *Proceedings of the 4th International Symposium on Image and Signal Processing and Analysis*, Zagreb, Croatia, 2005, 64–69. https://doi.org/10.1109/ispa.2005.195385
9. A. Doucet, N. Freitas, N. Gordon, *Sequential Monte Carlo methods in practice*, New York, NY: Springer, 2001. https://doi.org/10.1007/978-1-4757-3437-9
10. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial networks, *Commun. ACM*, **63** (2020), 139–144. https://doi.org/10.1145/3422622
11. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, et al., Generative adversarial nets, In: *Advances in Neural Information Processing Systems*, NIPS 2014, Volume 27, Curran Associates, Inc., 2014.
12. J. D. Hol, T. B. Schon, F. Gustafsson, On resampling algorithms for particle filters, *2006 IEEE Nonlinear Statistical Signal Processing Workshop*, 2006, 79–82. https://doi.org/10.1109/nsspw.2006.4378824
13. K. Van Koevering, J. Kleinberg, How random is random? Evaluating the randomness and humaness of LLMs' coin flips, *arXiv*, 2024. https://doi.org/10.48550/arXiv.2406.00092
14. M. A. Kouritzin, Convergence rates for residual branching particle filters, *J. Math. Anal. Appl.*, **449** (2017), 1053–1093. https://doi.org/10.1016/j.jmaa.2016.12.046

15. M. A. Kouritzin, Residual and stratified branching particle filters, *Comput. Stat. Data Anal.*, **111** (2017), 145–165. https://doi.org/10.1016/j.csda.2017.02.003

16. M. A. Kouritzin, Markov observation models and deepfakes, *Mathematics*, **13** (2025), 2128. https://doi.org/10.3390/math13132128

17. M. A. Kouritzin, F. Newton, S. Orsten, D. C. Wilson, On detecting fake coin flip sequences, *Inst. Math. Stat. (IMS) Collect.*, **4** (2008), 107–122. https://doi.org/10.1214/074921708000000336

18. M. A. Kouritzin, F. Newton, B. Wu, A flexible, real-time algorithm for simulating correlated random fields and its properties, *J. Math. Stat.*, **13** (2017), 197–208. https://doi.org/10.3844/jmssp.2017.197.208

19. L. Li, J. Bao, T. Zhang, H. Yang, D. Chen, F. Wen, et al., Face X-ray for more general face forgery detection, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, 5000–5009. https://doi.org/10.1109/cvpr42600.2020.00505

20. Y. Li, X. Yang, P. Sun, H. Qi, S. Lyu, Celeb-DF: A large-scale challenging dataset for deepFake forensics, *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, WA, USA, 2020, 3204–3213. https://doi.org/10.1109/cvpr42600.2020.00327

21. L. M. Murray, A. Lee, P. E. Jacob, Parallel resampling in the particle filter, *J. Comput. Graph. Stat.*, **25** (2016), 789–805. https://doi.org/10.1080/10618600.2015.1062015

22. A. Odena, C. Olah, J. Shlens, Conditional image synthesis with auxiliary classifier GANs, *Proceedings of the 34th International Conference on Machine Learning*, 2017, 2642–2651.

23. W. Pieczynski, Pairwise markov chains, *IEEE Trans. Pattern Anal. Mach. Intell.*, **25** (2003), 634–639. https://doi.org/10.1109/tpami.2003.1195998

24. L. R. Rabiner, A tutorial on hidden Markov models and selected applications in speech recognition, *Proc. IEEE*, **77** (1989), 257–286. https://doi.org/10.1109/5.18626

25. A. Renelle, S. Budgett, E. Chernoff, Making heads and tails of generation loss: a timeless tale of folk randomness, *Proceedings of the Eleventh International Conference on Teaching Statistics*, 2023.

26. P. Révész, Strong theorems on coin tossing, *Proceedings of the International Congress of Mathematicians (Helsinki, 1978)*, **2** (1980), 749–754.

27. A. Rossler, D. Cozzolino, L. Verdoliva, C. Riess, J. Thies, M. Niessner, FaceForensics++: Learning to detect manipulated facial images, *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, Seoul, Korea (South), 2019, 1–11. https://doi.org/10.1109/iccv.2019.00009

28. M. F. Schilling, The longest run of heads, *Coll. Math. J.*, **21** (1990), 196–207. https://doi.org/10.1080/07468342.1990.11973306

29. T. Taulli, The impact on major industries, In: T. Taulli, *Generative AI: How ChatGPT and other AI tools will revolutionize business*, Berkeley, CA: Apress, 2023, 175–188. https://doi.org/10.1007/978-1-4842-9367-6_8

30. T. Wang, X. Liao, K. P. Chow, X. Lin, Y. Wang, Deepfake detection: A comprehensive survey from the reliability perspective, *ACM Comput. Surv.*, **57** (2024), 1–35. https://doi.org/10.1145/3699710

31. P. A. Warren, U. Gostoli, G. D. Farmer, W. El-Deredy, U. Hahn, A re-examination of "bias" in human randomness perception, *J. Exp. Psychol.: Human Percept. Perform.*, **44** (2018), 663–680. https://doi.org/10.1037/xhp0000462

32. M. Westerlund, The emergence of deepfake technology: a review, *Technol. Innov. Manag. Rev.*, **9** (2019), 39–52. https://doi.org/10.22215/timreview/1282

33. J. Xiong, *An introduction to stochastic filtering theory*, Oxford: Oxford University Press, 2023. https://doi.org/10.1093/oso/9780199219704.001.0001

34. K. Yang, W. Y. Lin, M. Barman, F. Condessa, Z. Kolter, Defending multimodal fusion models against single-source adversaries, *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Nashville, TN, USA, 2021, 3339–3348, https://doi.org/10.1109/cvpr46437.2021.00335

35. X. Zhu, Y. Wang, E. Cambria, I. Rida, J. S. López, L. Cui, et al., RMER-DT: Robust multimodal emotion recognition in conversational contexts based on diffusion and transformers, *Inf. Fusion*, **123** (2025), 103268. https://doi.org/10.1016/j.inffus.2025.103268

AIMS Press