Applied Physics Letters

A perspective on deep neural network-based detection for multilayer magnetic recording

Cite as: Appl. Phys. Lett. **119**, 010502 (2021); https://doi.org/10.1063/5.0051085
Submitted: 22 March 2021 . Accepted: 12 June 2021 . Published Online: 07 July 2021

Ahmed Aboutaleb, Amirhossein Sayyafan, Krishnamoorthy Sivakumar, Benjamin Belzer, Simon Greaves, Kheong Sann Chan, Roger Wood, et al.

View Online Export Citation CrossMark

ARTICLES YOU MAY BE INTERESTED IN

Perspective—On the thermodynamics of perfect unconditional security
Applied Physics Letters **119**, 010501 (2021); https://doi.org/10.1063/5.0057764

Doping of organic semiconductors: Insights from EPR spectroscopy
Applied Physics Letters **119**, 010503 (2021); https://doi.org/10.1063/5.0054685

Electrical tuning of the spin-orbit interaction in nanowire by transparent ZnO gate grown by atomic layer deposition
Applied Physics Letters **119**, 013102 (2021); https://doi.org/10.1063/5.0051281

Applied Physics Letters PERSPECTIVE scitation.org/journal/apl

A perspective on deep neural network-based detection for multilayer magnetic recording

Cite as: Appl. Phys. Lett. **119**, 010502 (2021); doi: 10.1063/5.0051085
Submitted: 22 March 2021 · Accepted: 12 June 2021 ·
Published Online: 7 July 2021

Ahmed Aboutaleb,[1] Amirhossein Sayyafan,[1] Krishnamoorthy Sivakumar,[1] Benjamin Belzer,[1,a)] Simon Greaves,[2] Kheong Sann Chan,[3] and Roger Wood[1]

AFFILIATIONS

[1]School of Electrical Engineering and Computer Science, Washington State University, Pullman, Washington 99164, USA
[2]Research Institute of Electrical Communication (RIEC), Tohoku University, Sendai 980-8577, Japan
[3]Department of Electrical and Electronic Engineering, University of Nottingham Malaysia, Semenyih 43500, Malaysia

[a)]Author to whom correspondence should be addressed: belzer@wsu.edu

ABSTRACT

This paper describes challenges, solutions, and prospects for data recovery in multilayer magnetic recording (MLMR)—the vertical stacking of magnetic media layers to increase information storage density. To this end, the channel model for MLMR is discussed. Data recovery is described in terms of the readback stage followed by equalization and then detection. We illustrate how deep neural networks (DNNs) can be used to design systems for equalization and detection for MLMR. We show that such DNN-based systems outperform the conventional baseline and provide a good trade-off between complexity and performance. To achieve additional density gains, several prospective methods are discussed. On a physical level, the selective reading of tracks on different layers can be achieved by resonant reading. Resonant reading promises reduced interference from different layers, enabling higher storage densities. Regarding the signal processing, DNNs can be used to estimate the media noise and iteratively exchange soft-bit information with the decoder. Also, to ameliorate partial erasures, an auto-encoder-based system is proposed as a modulation coding scheme.

I. INTRODUCTION

Since the introduction of commercial hard disk drives (HDDs) in 1957, the areal density has increased from 2000 bit/in.^{2} to more than 1 Terabit/in.^{2} in 2020. However, the bits in an HDD are currently stored on the two-dimensional (2D) surface of the magnetic medium. If data could be stored in multiple, discrete magnetic layers, the storage capacity could be increased significantly.

Modern HDDs store bits as either positive or negative magnetizations of magnetic grains oriented perpendicular to the disk surface and support a density of more than 1 Terabit/in.^{2}, with about 10 magnetic grains per bit and about 10 Teragrains/in.^{2}. In recent years, increasing the information density by shrinking the average grain size has run up against the superparamagnetic limit in which random thermal variations flip grain magnetizations, resulting in the "media trilemma": reducing the bit cell size without reducing the average grain size leads to fewer grains per bit, thereby degrading the media signal-to-noise ratio (SNR); reducing the average grain size leads to thermal grain magnetization flipping which degrades the stored information's longevity; alleviating thermal flipping by increasing the grains' anisotropy K_u makes them harder to write to the point that there will be insufficient field from the write head to flip the grains.

The superparamagnetic limit has spurred research on new technologies for increasing HDD information densities. Most HDDs write data-bits on each track independently. The track-pitch (TP), i.e., the track center-to-center distance, must be large enough to reduce intertrack interference (ITI), which occurs when the read head picks up magnetic signals from adjacent tracks, to an acceptably low level. Down-track intersymbol interference (ISI) is reduced by using the Viterbi algorithm (VA) on the state trellis that is defined by the downtrack interference between the bits.[1]

Two dimensional magnetic recording (TDMR), proposed in Ref. 2, increases density by decreasing the TP and writing and reading bits in several tracks simultaneously. In this case, the readback system must deal with interference in two dimensions (down-track ISI and cross-track ITI). The higher data density reduces the number of grains per bit. This not only reduces the available SNR, but can also cause further degradation from interactions between data on the closely-spaced tracks. These interactions include both nonlinear signal distortion and complex magnetostatic interactions that increase signal-dependent

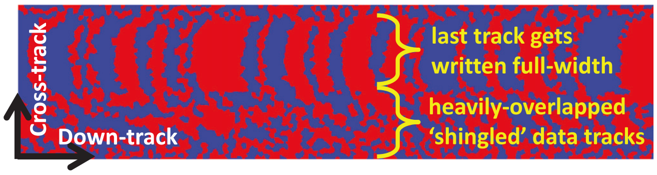

FIG. 1. A simulated TDMR magnetization signal written on the disk surface. The blue and red stripes represent negatively and positively magnetized bits. Bit regions are curved stripes due to the relative orientation of the corner write head. To increase density, bits overlay one another both cross-track and down-track in a technique called "shingled writing." Reproduced with permission from J. Shen, A. Aboutaleb, K. Sivakumar, B. J. Belzer, K. S. Chan, and A. James, IEEE Trans. Magn. **56**, 3100212 (2020). Copyright 2020 IEEE.[13]

media noise.[3] Trellis-based detection methods have been proposed for TDMR, e.g., Refs. 4–9, but most suffer from an explosion in the number of trellis states when more than two tracks are read simultaneously.

Figure 1 shows a capture of five tracks of a written TDMR magnetization signal produced by a micromagnetic-based simulation. The blue and red stripes represent negatively and positively magnetized bits, corresponding, e.g., to binary 0 and 1 values. Bit regions are curved stripes due to the relative orientation of the corner write head. The lower portion of the image shows four shingled (heavily overlapped) tracks typical of the TDMR data that must be recovered. The upper portion of the picture shows the last track written and reveals the full width of the write head.

Figure 1 motivates considering TDMR bit detection as an image classification problem, leading to recent work on deep neural network (DNN) detectors for TDMR. DNNs, developed in 2006, have achieved remarkable results in image classification and image understanding.[10] Recently, a number of papers (e.g., Refs. 11–18) have applied DNN-based signal processing to TDMR detection, with significant success: Ref. 17 achieves an areal information density of 3.88 Tb/in.^{2} by processing three tracks of micromagnetic-simulated TDMR waveforms with about 2.9 grains per bit and 11 Teragrains/in.^{2}. The work in Ref. 19 and that in references therein are among the first to propose nonlinear equalization using neural networks for 1DMR channels. In Refs. 11–14, the neural network equalizer is extended to TDMR. Recent studies have proposed using a convolutional neural network (CNN) detector to estimate the written bits without the equalization stage for TDMR[13,18] and MLMR.[20] In Ref. 17, the DNN is used as a media noise predictor that can be integrated with conventional and CNN-based equalization and detection sub-systems.

Recent encouraging studies[20–26] propose multilayer magnetic recording (MLMR): vertical stacking of an additional magnetic media layer to a TDMR system to achieve further density gains. Figure 2 illustrates a two-layer recording structure. The lower layer is farther apart from the read head compared with the upper layer, resulting in a weaker signal from the lower layer. To compensate for this, the bit area on the lower layer is four times larger than that on the upper layer. In general, the ratio of the number of bits on the upper layer to that on the lower layer is a system parameter. It is expected that this ratio can be tuned to maximize the total density gain, while maintaining acceptable error rates. The respective layers use different bit sizes and can be written at different frequencies using microwave assisted magnetic recording (MAMR). Using a realistic grain switching probability (GSP) model to generate waveforms in a two-layer MLMR system as in Refs. 22–24, the coauthors of the present paper investigated DNN based methods for equalization and detection for two-layer MLMR in Ref. 20 and reported significant density gains due to the additional lower layer. MLMR requires *joint signal separation and equalization*: the readers lie just above the disk surface and hence receive a superposition of signals from the upper and lower layers, and the received signal now suffers from 3D-ISI due to per layer ITI and downtrack ISI, plus inter-layer inference (ILI). These problems are somewhat ameliorated by the interleaved MLMR proposed in Refs. 27 and 28, wherein upper layer tracks only partially overlay lower layer tracks, but at the cost of less potential density gain than the MLMR in Ref. 20, wherein two upper layer tracks completely overlay one lower layer track.

II. MULTILEVEL MAGNETIC RECORDING: PHYSICS AND SIMULATION METHODS

A. Multilevel magnetic recording

Over the years, there have been several proposals to realize multilayer magnetic recording. An early example is that of Yuan *et al.*,[29] who suggested a double layer system combining media with longitudinal and perpendicular anisotropy.

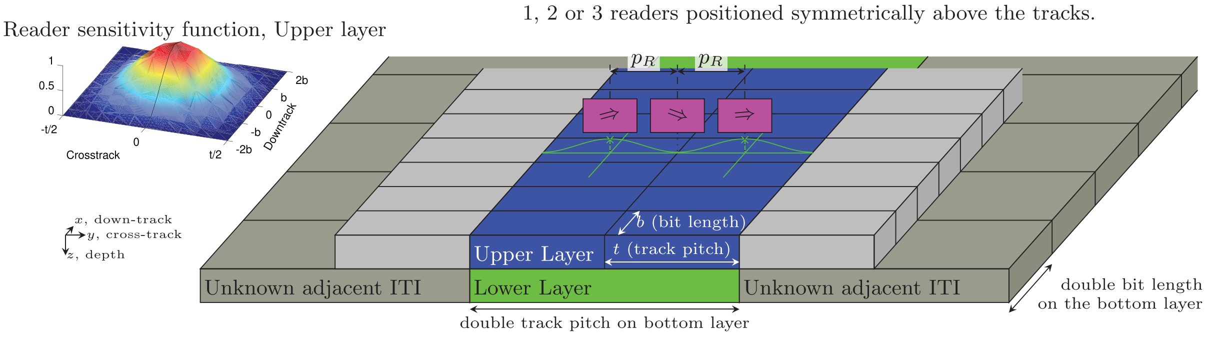

FIG. 2. A two-layer magnetic recording structure. The 2D reader sensitivity function gives the 2D response of the read head. Bits on the lower layer are written at quarter density compared with the upper layer, as indicated by the double bit length and track pitch. Although omitted here, the two layer recording structure includes a non-magnetic material in-between layers and a soft magnetic underlayer underneath the lower layer, which are shown in Fig. 3. Reproduced with permission from K. S. Chan, A. Aboutaleb, K. Sivakumar, B. Belzer, R. Wood, and S. Rahardja, IEEE Trans. Magn. **55**, 6700605 (2019). Copyright 2019 IEEE.[11]

Following the introduction of perpendicular media, it was suggested that media with two,[30] three,[31] or more[32] layers, each with a different coercivity, could be used. Data could be written on one or more layers by modulating the strength of the field produced by the write head.

However, the field generated by a write head varies non-linearly with write current and is difficult to control precisely. Given this, energy-assisted recording appears more suited to multilayer magnetic recording. Consider, for example, heat-assisted magnetic recording (HAMR).[33,34] If the recording layers in a HAMR system have different Curie temperatures, then the maximum temperature during recording can be used to determine which layers are written.[35,36]

One problem with these methods is that, in order to write on the layer with higher coercivity or higher Curie temperature, data on the layer with lower coercivity or lower Curie temperature are erased. Thus, data must be written on the layers in a particular order.

Microwave-assisted magnetic recording (MAMR)[37] is an alternative type of energy-assisted recording. In a MAMR system, a high frequency (HF) magnetic field oscillating at, or near to, the ferromagnetic resonance frequency of a recording layer can lower the switching field of that layer. It follows that if two recording layers have different ferromagnetic resonance frequencies, selective recording on either of the two layers becomes possible by varying the frequency of the HF field.[38–40]

Figure 3 shows a schematic of a multilevel MAMR system. The HF field is generated by a spin torque oscillator (STO) located between the main pole and the trailing shield of the write head. The inset shows the results of a simulation in which the switching probability of the upper and lower recording layers was calculated as a function of the HF field frequency. Depending on the HF field frequency, data could be written on either the upper or the lower recording layer.

B. Grain switching probability simulation of MLMR

This subsection explains the micromagnetic based grain switching probability simulation of MLMR employed in Ref. 20 and in the present article. A cross-track view of the two-layer system considered is shown in Fig. 4. The bit sequences written on the upper left and right tracks are denoted by $\mathbf{a}_{2,L}$ and $\mathbf{a}_{2,R}$, respectively, and the bit sequence on the lower track is denoted by $\mathbf{a}_1$. The reading sequences $\mathbf{r}_L$, $\mathbf{r}_C$, and $\mathbf{r}_R$ are observed above the left, center, and right tracks, respectively. The three reading sequences can be simultaneously obtained by a novel three read head configuration. The GSP model in Refs. 22 and 23 is used to simulate the write process for a two-layer system.

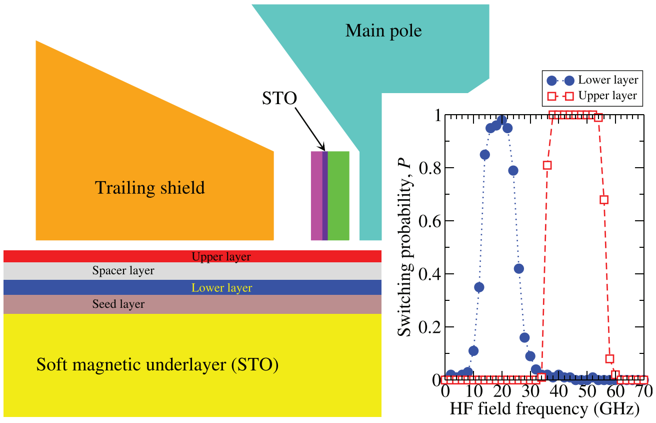

FIG. 3. Schematic of a two layer MAMR system with two recording layers. Inset: Switching probability of the upper and lower recording layers vs HF field frequency.

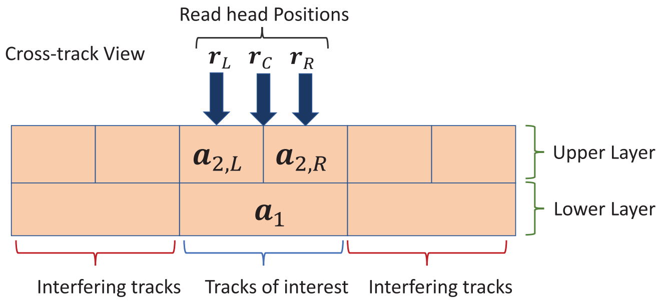

FIG. 4. Cross-track view of a two-layer MLMR system. The bit sequences on the upper left, upper right, and lower tracks are denoted by $\mathbf{a}_{2,L}$, $\mathbf{a}_{2,R}$, and $\mathbf{a}_1$, respectively. Reading sequences taken at read head positions left, center, and right are denoted by $\mathbf{r}_L$, $\mathbf{r}_C$, and $\mathbf{r}_R$, respectively. The non-magnetic spacing between layers and the soft-magnetic underlayer (shown in Fig. 3) are omitted here for simplicity. Reproduced with permission from A. Aboutaleb, A. Sayyafan, K. Sivakumar, B. Belzer, S. Greaves, K. S. Chan, and R. Wood, IEEE Trans. Magn. **57**, 3101012 (2021). Copyright 2021 IEEE.[20]

The write process involves magnetizing grains within a bit cell according to the desired bit polarity. To simulate the stochastic write process, the GSP model provides the probability that a grain will switch in the direction of the head field. This switching probability is a function of the applied field strength, the distance between the grain and the write head, and the magnetization polarities of nearby grains.[20]

Importantly, the magnetostatic interactions with nearby grains in the same recording layer play a key role in determining the polarity of a target grain. For instance, if nearby grains are magnetized with the same polarity as the target grain's initial polarity, the magnetostatic field from the nearby grains will assist switching of the target grain, and in some cases unwanted switching may occur.

Conversely, if the net magnetization of nearby grains is in the opposite direction to the target grain, the switching probability of the target grain will be decreased. In such a case, the target grain may not switch at all. Thus, magnetostatic interactions can result in an increase in media noise and partial erasures.

Magnetostatic interactions between recording layers should also be considered. Magnetostatic interactions between vertically adjacent grains magnetized in the same direction will impede switching of the magnetization and vice versa.

Let $\mathbf{a}'_k$ denote the actual granular magnetization pattern for layer k after the write process for the intended pattern $\mathbf{a}_k$. Then the likelihood $\mathcal{P}\{\mathbf{a}'_k|\mathbf{a}_k\}$ of obtaining $\mathbf{a}'_k$ given $\mathbf{a}_k$ is implicitly modeled by the GSP model. The GSP model is trained to reproduce the data based on micromagnetic simulations of the system, which are computationally expensive. Hence, the GSP model provides an accurate yet computationally efficient method for generating data.[20,22]

The read process is the first stage for recovering the written bits. The read heads are positioned above the upper layer. They measure the magnetic fields arising from the granular magnetization pattern corresponding to the written bits. The signal associated with bit n depends on the polarity of n and the polarities of neighboring bits.

Such neighboring bits reside in the down-track and the cross-track directions, resulting in 2D ISI/ITI. For MLMR, ILI from having multiple layers also affects the read signal. This read signal is also called the readback signal. Let $h_k[i,j]$ represent the 2D-ISI/ITI response for layer k, and $r_m[n]$ represent the discrete-time readback signal, measured at the down-track bit position n and the cross-track bit position m. For the two-layer recording system, $r_m[n]$ is given by

$$r_m[n] = \sum_{k=1}^{2}\sum_{i,j} h_k[i,j]a'_k[m-i,n-j] + n_m[n], \tag{1}$$

where $n_m[n]$ is additive white Gaussian noise (AWGN) with zero-mean and variance σ_e^2 to model the reader electronics noise.

III. DEEP NEURAL NETWORK BASED DETECTION AND EQUALIZATION FOR MLMR

Traditional equalization is performed using a 2D linear minimum mean squared error (2D-LMMSE) equalizer. Typically, the equalizer is realized as a finite impulse response (FIR) filter. Equalization is then followed by maximum likelihood (ML) detection using the VA.[1] The equalizer's output is approximated as a convolution of the input binary bit sequence with a target filter. The 2D partial response (PR) target is designed to have fewer taps than the natural ISI/ITI span of the channel response. In Ref. 26, a 1024-state VA is developed for two-layer recording, where the channel model does not include partial erasures and transition noise. The baseline system consisting of the 2D-LMMSE followed by the 1024-state VA is detailed in Ref. 20.

Two approaches for using CNNs for equalization and detection are now discussed.

A. CNN equalizer–separator–SOVA system

Appropriately designed CNNs can be used to separate and partially equalize the data sequences from the reading sequences. The output of the CNN approximates a noise-free PR signal. To obtain soft-bit estimates of the written binary sequence, the output of the CNN is then fed to a regular soft-output VA (SOVA).[41] The soft-bit estimate refers to a reliability measure associated with the bit estimate, as opposed to hard-bit estimation where the bit is detected as 0 or 1 without any reliability information. Such soft-bit estimates are then used by an irregular repeat accumulate (IRA) decoder to recover the information bits.

The proposed CNN equalizer-separator-SOVA system uses CNNs for equalization and separation, which are followed by three 1D-2^S-state Viterbi detectors as shown in Fig. 5, where S is the order of the PR target. The CNNs accept raw readings $\mathbf{r}_L$, $\mathbf{r}_C$, and $\mathbf{r}_R$ as input and output separated 1D equalized sequences $\hat{\mathbf{s}}_1$, $\hat{\mathbf{s}}_{2,L}$, and $\hat{\mathbf{s}}_{2,R}$. The equalization follows target $\mathbf{g}_k$, for layer k, where $k = 1, 2$. The target $\mathbf{g}_k$ is tuned to minimize the mean-squared error (MSE) given fixed CNN weights, per the following procedure. The noise-free PR signals per track of interest are given by

$$\mathbf{s}_1 = \mathbf{g}_1 * \mathbf{a}_1, \tag{2}$$

$$\mathbf{s}_{2,L} = \mathbf{g}_2 * \mathbf{a}_{2,L}, \tag{3}$$

$$\mathbf{s}_{2,R} = \mathbf{g}_2 * \mathbf{a}_{2,R}, \tag{4}$$

where $\mathbf{g}_1$ and $\mathbf{g}_2$ are $(S+1)$-tap targets. In general, $\mathbf{g}_1$ and $\mathbf{g}_2$ need not have the same number of taps. Two CNNs, one CNN per layer, are trained to equalize and separate the signals corresponding to each layer. For given targets, the CNNs are trained to minimize the sample MSE between their outputs and the noiseless PR signals. The cost functions to minimize are given by

$$J_{\mathrm{MSE},1} = \frac{2}{N}\sum_{n=0}^{N/2-1}(\hat{s}_1[n] - s_1[n])^2, \tag{5}$$

$$J_{\mathrm{MSE},2} = \frac{1}{2N}\sum_{m\in\{L,R\}}\sum_{n=0}^{N-1}(\hat{s}_{2,m}[n] - s_{2,m}[n])^2 \tag{6}$$

for the lower and upper layers, respectively. The optimization is performed numerically using a variant of stochastic gradient descent called the Adam optimizer.[42] The convergence of the CNN training is indicated by negligible reductions in the cost functions as the training iterations continue. After the convergence of the CNN training, the CNN weights are fixed. The CNN outputs are used to adapt the targets to further minimize the MSE. The target optimization problem is given by (for $k = 1, 2$)

$$\underset{\mathbf{g}_k}{\text{minimize}} \quad J_{\mathrm{MSE},k} \tag{7a}$$

$$\text{subject to} \quad \mathbf{u}_k^T\mathbf{g}_k = 1, \quad c_{k,\min} \le ||\mathbf{g}_k||_2^2 \le c_{k,\max}, \tag{7b}$$

where $\mathbf{u}_k$ imposes a monic constraint on $\mathbf{g}_k$, e.g., dictating that the center tap is one, and the square Euclidean norm of $\mathbf{g}_k$ is bounded in the interval $[c_{k,\min}, c_{k,\max}]$ with $0 \le c_{k,\min} \le c_{k,\max} < \infty$. Following the targets' optimization, the CNN training is continued with the new

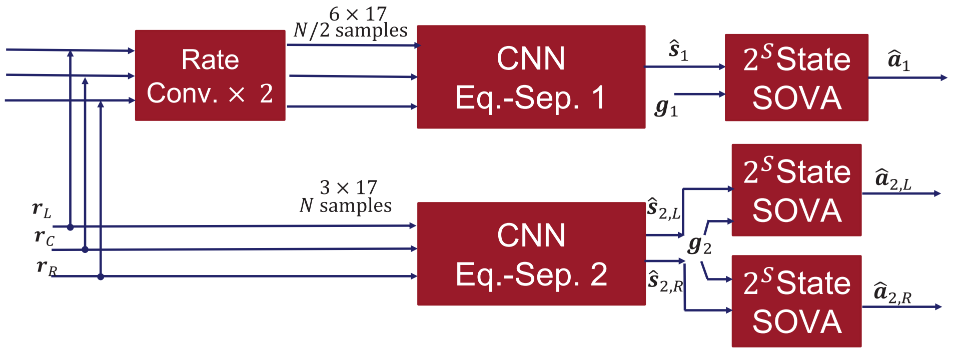

FIG. 5. The CNNs accept readings from the channel and output equalized and separated streams per track of interest. A regular 1D SOVA follows for soft-bit detection. Reproduced with permission from A. Aboutaleb, A. Sayyafan, K. Sivakumar, B. Belzer, S. Greaves, K. S. Chan, and R. Wood, IEEE Trans. Magn. **57**, 3101012 (2021). Copyright 2021 IEEE.[20] CNN equalizer– separator–SOVA system.

targets. This iterative process repeats until no substantial reductions in MSE are observed.

Figure 6 shows an example CNN architecture for the CNN equalizer-separator system. The inputs to the CNN equalizer are readings $\mathbf{r}_L$, $\mathbf{r}_C$, and $\mathbf{r}_R$ over sliding windows of size 3 × 17 and 6 × 17 for the upper and lower layers, respectively. The size of the sliding window is selected based on the approximate span of the ISI/ITI. To maintain a 17-bit down-track footprint, the additional readings are multiplexed across track for the lower layer's CNN. The type of convolution operation performed is "same," which means that the input is padded with zeros at the beginning of the input sequence such that the size of the output is the same as the size of the input. Hence, for the upper layer's CNN, the output of each convolutional layer is size $3 \times 17 \times D$, where D is the number of groups of convolutional filters; D is 10, 8, 6, and 4 for the first, second, third, and fourth convolutional layers, respectively. The number of filters per group at the current layer is equal to the number of groups of convolutional filters in its preceding layer. For example, the first layer consists of 10 groups of convolutional filters. Each convolutional filter is a 2D FIR filter of size 3 × 11. This allows the filter to capture correlations in the input samples spanning a 3 × 11 window of readback samples. To reduce the chance of over-fitting, a dropout layer is applied at the output of the first convolutional layer.[43]

The rectified linear unit (ReLU) is used as the non-linear activation function applied on the outputs of intermediate layers. The ReLU activation function is defined by the element-wise relation $f(x) = \max(0, x)$. For the upper layer's CNN, the output of the last convolutional layer is flattened to a column vector of size 204 × 1 before being fed to the last layer, which is a fully connected (FC) layer comprised of a matrix of size 204 × 2 and two bias variables. The outputs of the FC layer are the estimates of the equalized and separated signals $\hat{s}_{2,L}[n]$ and $\hat{s}_{2,R}[n]$ for the upper layer. For the lower layer's CNN, the output of the last convolutional layer is of size 408 × 1. The output of the FC layer is computed as an affine combination of the 408 variables to give the estimate $\hat{s}_1[n]$.

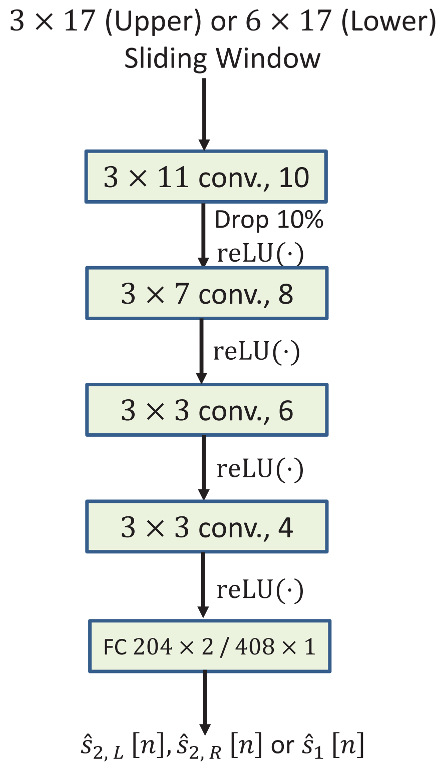

FIG. 6. CNN architecture for the equalizer–separator–VA system. Reproduced with permission from A. Aboutaleb, A. Sayyafan, K. Sivakumar, B. Belzer, S. Greaves, K. S. Chan, and R. Wood, IEEE Trans. Magn. **57**, 3101012 (2021). Copyright 2021 IEEE[20].

B. CNN detector system

The CNN detector system shown in Fig. 7 uses CNNs to detect bits directly from raw readings, rather than as an intermediary system before the SOVA. Hence, this system subsumes the equalization and detection sub-systems within one system. The motivation is that the CNNs in this system can now be trained to directly minimize an objective function that more closely reflects the bit error rate (BER) metric than the MSE objective. Indeed, the cross-entropy (CE) loss, computed between the soft estimate of the bits and the true bits, can result in lower BERs than the MSE loss.[14]

Furthermore, the bit detection problem can be viewed as an image classification problem. The readback raw readings contained within a sliding window can be used to detect bits on the upper or lower layers. In this view, the samples contained within each interval of the sliding window constitute an image whose correct classification label is the true bit at the center of the window in the down-track direction. CNNs have been successful at accurately classifying images when appropriately trained.[44] Thus, the CNN provides a promising method for bit detection over realistic digital storage channels when training data are available, but the channel model is difficult to characterize.

Consider estimating the nth bit $a_L[n]$, and let $\hat{a}_L[n]$ represent its estimate. Let the indicator function $\mathbb{1}_{a_L[n]=i} = 1$ if $a_L[n] = i$, $i = 0, 1$, and zero otherwise. Then the CE loss is defined as

$$\begin{aligned}\mathcal{H}\{a_L[n], \hat{a}_L[n]\} = &-\mathbb{1}_{a_L[n]=0} \log\left(\Pr\{\hat{a}_L[n] = 0\}\right) \\ &- \mathbb{1}_{a_L[n]=1} \log\left(1 - \Pr\{\hat{a}_L[n] = 0\}\right). \qquad (8)\end{aligned}$$

The objective function to minimize is the average CE loss $J_{\mathrm{CE}} = (1/N)\sum_{n=0}^{N-1} \mathcal{H}\{a_L[n], \hat{a}_L[n]\}$ computed over a length-N mini-batch. During training, the backpropagation algorithm with stochastic

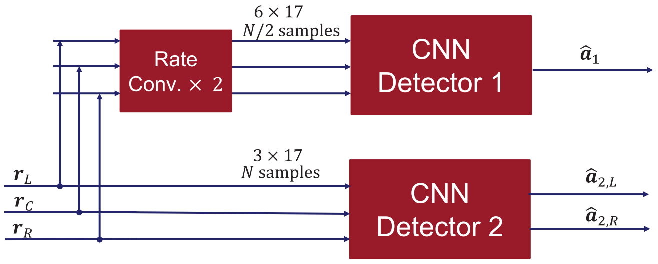

FIG. 7. CNN detector system. The CNNs detect bits directly from raw readings without an intermediary equalization sub-system. Reproduced with permission from A. Aboutaleb, A. Sayyafan, K. Sivakumar, B. Belzer, S. Greaves, K. S. Chan, and R. Wood, IEEE Trans. Magn. **57**, 3101012 (2021). Copyright 2021 IEEE[20].

gradient descent (cf. Ref. 44, Sec. 8.1.3) is used to adapt the learnable parameters of the system to minimize the objective function.

Figure 8 shows the architectures of two CNNs for detection, named D1 and D2. D1 entails a higher implementation complexity than D2 but achieves improvement in the detection BER. It has been shown in Ref. 20 that CNNs readily allow for a performance-complexity trade-off.

C. Main results

The CNN-based systems were tested on realistic data generated by the GSP model, which is trained on micromagnetic simulations. The generated data consist of 100 blocks that were divided into 60 training blocks, 20 validation blocks, and 20 testing blocks. Each block contains 82, 412 bits per track for the upper layer and 41, 206 bits per track for the lower layer. For training the CNN equalizer-separator, a mini-batch of size 100 samples is used in the first iteration with the target solver. The mini-batch size is increased to 500 samples in subsequent iterations. Training the CNN detector D1 used a mini-batch size of 100 samples; this was increased to 1000 samples in subsequent iterations as recommended in Ref. 45. Training CNN detector D2 used a fixed mini-batch size of 1000 samples. For all CNNs, the adaptive learning rate starts at 10^{-3} and is decreased to 10^{-5} in the final

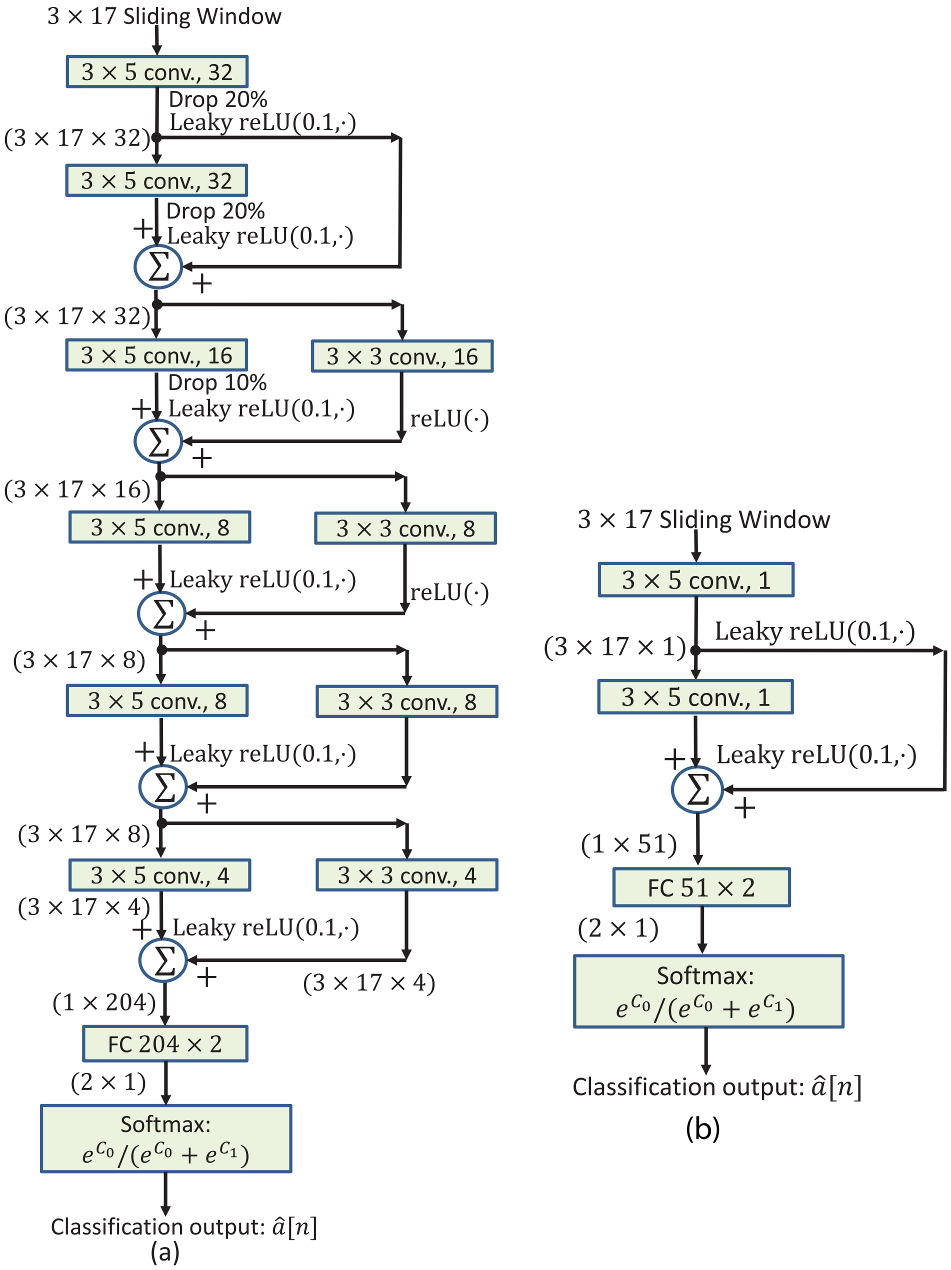

FIG. 8. Architectures of the CNN detector system. In (a), CNN detector D1 achieves a detection BER of 6.610%. In (b), the lower complexity CNN detector D2 achieves a detection BER of 7.304%. Reproduced with permission from A. Aboutaleb, A. Sayyafan, K. Sivakumar, B. Belzer, S. Greaves, K. S. Chan, and R. Wood, IEEE Trans. Magn. **57**, 3101012 (2021). Copyright 2021 IEEE.[20]

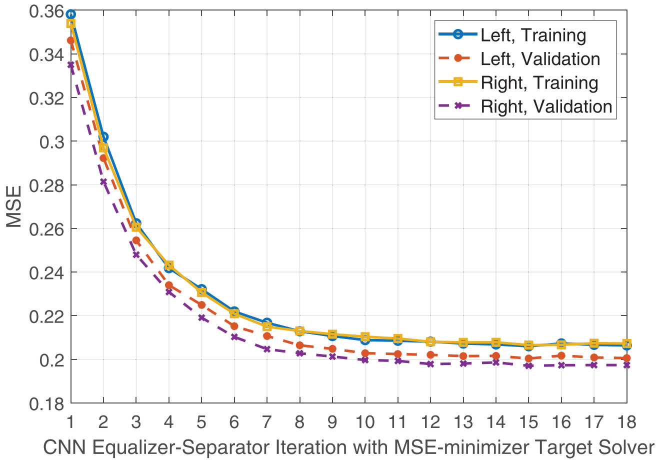

FIG. 9. MSE vs iteration between the CNN equalizer-separator and MSE-minimizer target solver for the left and right tracks on upper layer. The MSE converges to a minimum that satisfies the energy and monic constraints on the PR target.

few training iterations. The Adam optimizer is used during backpropagation.[42]

As a baseline comparison, we simulate the performance of the conventional system consisting of a 2D-LMMSE equalizer followed by a VA detector (2D-LMMSE-VA). For the baseline, the VA used is the 1024-state VA developed in Ref. 26 for a two-layer magnetic recording channel without jitter noise and partial erasures.

Figure 9 illustrates the reduction and convergence of the MSE in the iteration between the CNN equalizer-separator and the MSE target solver. The CNN equalizer-separator system output is fed to the SOVA detectors for estimating the written bits from the equalized and separated waveforms. The BER is computed by comparing the estimates with the true written bits.

Figure 10 shows the learning curve for training the CNN detector architectures D1 and D2. The training accuracy is computed using a random mini-batch, whereas the validation accuracy is computed using all the validation data.

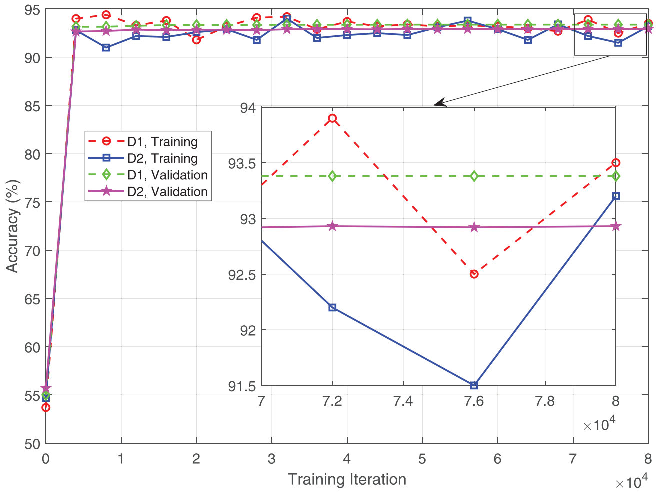

FIG. 10. Accuracy of bit detection vs the number of training iterations for CNN detectors D1 and D2.

Table I summarizes the detector BERs achieved by the CNN and baseline systems. The CNN-based systems outperform the conventional 2D-LMMSE-VA system in terms of BER. The lowest BER is achieved by a CNN detector system.

The CNN detector system outperforms the CNN equalizer-separator-SOVA system because of the training criteria and the presence of media noise. The CNN detector system is trained to minimize the CE loss, which corresponds to lower BERs and improved soft information.[14] In contrast, the CNN equalizer-separator-SOVA system is trained to minimize the MSE loss under the monic constraint. Also, the SOVA is optimal for linear ISI channels with AWGN. Due to the presence of data-dependent media noise, the ISI is generally nonlinear in the written bits, and the noise is correlated and data-dependent. Hence, the SOVA is suboptimal.

Several CNN architectures were explored with varying implementation complexity requirements. Overall, the CNN can provide a good trade-off between performance and complexity. This was shown by a CNN detector that requires lower complexity than the 2D-LMMSE-VA system while achieving lower BERs. More precisely, this CNN detector architecture requires about 18% less multipliers per bit estimate while achieving a 45.6% lower detector BER than the 2D-LMMSE-VA system. Note that the architectures of S1, S2, D1, and D2 are given in Ref. 20, where they are referred to as S3, S6, D1, and D4, respectively.

To estimate an areal density gain from the detector BER, the log-likelihood ratios (LLRs) computed by the detectors are fed to an IRA decoder, which assumes that the transmitted bits are coded using an IRA code. The IRA code is a form of low-density parity check (LDPC) code that allows encoding in linear time while achieving the channel capacity for the binary erasure channel.[46] The code rate is adjusted such that the decoder's BER is less than 10^{-5}. Higher code rates correspond to higher information density achieved. We used IRA decoders that leverage coset decoding to process the LLRs output by the CNN equalizer-separator and CNN detector systems. The data bits written on the recording medium are randomly generated. In the coset decoding method, the decoder generates a valid codeword which agrees with the random written data bit pattern in the information bit positions. An XOR operation between the random written bit pattern and the generated codeword is done to identify parity bit positions where the

TABLE I. Detector BERs for the MLMR system. The complexity is represented as a factor of the number of multiplications per bit estimate needed by baseline 2D-LMMSE-1024-VA, which is 4935 multiplications. CNNs enable a trade-off between complexity and performance.

Method	Upper BER	Lower BER	Complexity
Channel BER	0.224 7	0.213 5	N/A
2D-LMMSE-1024-VA	0.133 5	0.181 2	1×
CNN Eq.-Sep. S1 2-VA	0.068 54	0.107 6	14×
CNN Eq.-Sep. S2 4-VA	0.070 71	0.114 6	1.36×
CNN detector D1	0.066 10	0.102 0	440×
CNN detector D2	0.073 04	0.139 9	0.82×

TABLE II. Achieved information rates by the CNN eq.-sep.-SOVA and CNN detector for the two-layer and one-layer systems. Two-layer recording offers about 16% gain in areal-density over conventional one-layer recording.

Method (CNN)	Two-layer rate	One-layer rate	Gain (%)
Eq.-Sep.-SOVA	0.833 9	0.715 7	16.51
Detector	0.868 8	0.711 6	16.20

two differ. This XOR result is used during the decoding process to correctly decode the random written data bits.

The proposed systems were tested on one-layer magnetic recording and two-layer magnetic recording structures. For the CNN equalizer-separator system, the information rates are 0.8339 and 0.7157 for the two-layer and one-layer systems, respectively. In comparison, the CNN detector systems achieves information rates of 0.8688 and 0.7116 for the two-layer and one-layer systems, respectively. The achieved density gains of MLMR over TDMR processing of the upper layer only (without interference from the lower layer) are summarized in Table II.

The CNN equalizer-separator architecture achieves a density gain of 16.51% for the two-layer structure over the one-layer structure. For the CNN detector, the overall information areal density achieved on the two-layer structure is about 16.20% higher than the density achieved on the one-layer structure. However, for the two-layer structure, the total density achieved by the CNN detector is 4.19% higher than the density achieved by the CNN equalizer-separator-SOVA system. The reason is that the CNN detector is trained on the CE loss, whereas the MSE loss is used for the CNN equalizer-separator. Minimizing the CE corresponds more directly to reducing the detector BERs and improving the soft-bit information, compared with minimizing the MSE. Initial experiments with AWGN added at 20 dB SNR in the two-layer system show a reduction of the density gain achieved by a CNN detector to 14.32%.

IV. FUTURE PROSPECTS FOR DNN DETECTION OF MLMR CHANNELS

A. Advances in multilevel recording technology

Advances in magnetic recording, especially the MAMR technology, which is most consistent with MLMR, will lead to higher recording densities. Notably, in Ref. 21 the readback process is designed to ameliorate ILI by resonant reading. Different layers are assigned different ferromagnetic resonance frequencies. This allows selective reading of bits from different layers with significant reductions in ILI. Since ILI is a limiting factor in the data recovery, such resonant reading enables higher information densities.

As ILI is appropriately mitigated, it is possible that MLMR architectures with a more equal area ratio between upper- and lower-layer bits will emerge as MAMR technologies advance. For example, as illustrated in Fig. 11, three upper-layer tracks could overlay two lower-layer tracks, with lower-layer bits being twice as long as upper-layer bits, such that each lower-layer bit would have the area of three upper-layer bits. Further advances in MAMR technology may allow including additional layers to the MLMR stack, with the proviso that each additional lower layer would require larger bit areas than layers above it to ensure enough SNR for reliable reading, and probably larger grain sizes as well, to enable writing of separate layers by exploiting the different grain resonant frequencies. For example, as shown in Fig. 11, a third layer could be added below the two above-described layers, with the third layer having only one track, and each third-layer bit having an area equal to that of two or four middle-layer bits. The trends of decreasing track pitch and decreasing bit lengths, such that there are fewer magnetic grains per bit, are also expected to continue. These advances in MLMR media and read/write technologies will create a need for higher throughput signal processing systems that are increasingly robust to magnetic media noise and to increasing levels of ISI, ITI, and ILI.

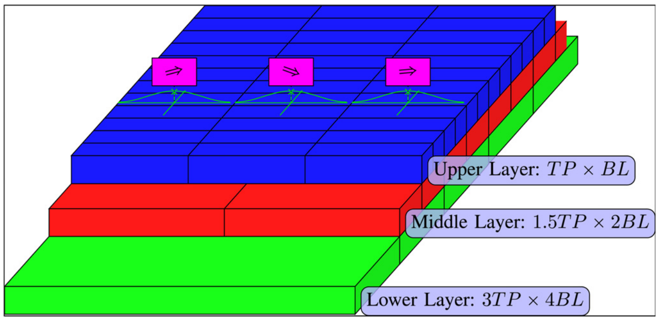
⇒
⇒
⇒
Upper Layer: $TP \times BL$
Middle Layer: $1.5TP \times 2BL$
Lower Layer: $3TP \times 4BL$

FIG. 11. A three-layer recording structure. The areas of bits on lower layers are larger compared with upper layers. Since the reader is placed above the upper layer, this allows high enough SNRs for the lower layers' signals to be captured by the reader.

Advances in MLMR will be enabled by increasingly sophisticated GSP simulations that model, for example, resonant reading, adding more layers, and the effects of inter-layer materials that might be added to reduce ILI or to facilitate better focusing of the write fields on each layer.

B. CNN media noise predictor for MLMR

CNNs have been used as noise predictors for 1D and 2D magnetic recording channels to achieve high information densities.[16,17] In such systems, the typical maximum *a posteriori* probability (MAP) detector assumes that the noise is AWGN for optimality in the MAP sense. In practice, the equalized signal can be given by

$$\hat{\mathbf{s}}_k = \mathbf{g}_k * \mathbf{a}_k + \mathbf{n}_{m,k} + \mathbf{n}_e, \qquad (9)$$

where $\mathbf{n}_{m,k}$ is the media noise for layer k, $\mathbf{g}_k$ is the 2D PR target for layer k, and $\mathbf{n}_e$ is the reader electronics noise modeled as a zero-mean Gaussian vector with independent components. To reduce the impact of the media noise, which is data-dependent and correlated, the CNN noise predictor estimates the media noise. The media noise estimate is then subtracted from the input to the MAP detector. The soft-bit decisions from the MAP detector can then be fed back to the CNN noise predictor to improve its estimation and the proceeding noise cancelation. Typically, the MAP detector is implemented using the Bahl, Cocke, Jelinek and Raviv (BCJR) algorithm,[47] which estimates soft information in the form of log-likelihood ratios (LLRs). This CNN-BCJR sub-system can iteratively exchange soft-bit information with a conventional LDPC decoder, which decodes the information bits.

CNN noise predictors have not been explored for MLMR. Figure 12 shows a block diagram of a proposed MLMR CNN noise predictor

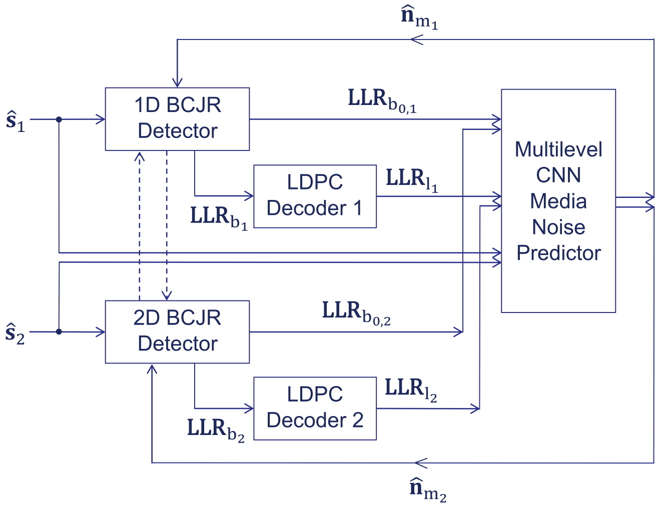

FIG. 12. MLMR turbo-detection system with multilevel CNN media noise predictor. The equalized and separated outputs from the CNN equalizer-separator are used as input for CNN media noise predictor and BCJR detectors. The LLRs from the BCJR are also fed to the CNN. An LDPC decoder exchanges LLRs with the BCJR. The LLRs from the decoder can also be fed to the CNN in a second iteration of the noise estimation followed by BCJR LLR estimation. Note that $\mathbf{s}_k$ can be a multi-dimensional array depending on the number of readers and the number of tracks that are simultaneously processed for layer k.

turbo-detection system. The inputs to the CNN noise predictor system are the equalized and separated sequences from the CNN equalizer-separator system discussed in Sec. III A. For the lower layer, $\hat{\mathbf{s}}_1$ is fed to a 1D BCJR. For the upper layer, $\hat{\mathbf{s}}_2$ is passed to a 2D BCJR since 2D ITI is expected to be more substantial on the upper layer than on the lower layer. The 1D and 2D BCJRs generate the LLR outputs $\mathbf{LLR}_{b_{0,1}}$ and $\mathbf{LLR}_{b_{0,2}}$ for the lower and upper layers, respectively.

For the first iteration of the turbo-detection system, a joint multilevel CNN noise predictor is provided with the equalized inputs and the LLRs from both layers to estimate media noise; using one CNN with inputs from both layers enables the CNN to account for inter-layer interactions in its media noise estimate. Then, the estimated media noise sequences $\hat{\mathbf{n}}_{m_1}$ and $\hat{\mathbf{n}}_{m_2}$ are used for the second pass of the BCJRs for the lower and upper layers, respectively. This allows each BCJR to improve the reliability of its LLRs. These new improved LLRs are denoted by $\mathbf{LLR}_{b_1}$ and $\mathbf{LLR}_{b_2}$. The improved LLRs are then fed to LDPC decoders. The decoders generate LLRs $\mathbf{LLR}_{l_1}$ and $\mathbf{LLR}_{l_2}$ for the lower and upper layers, respectively. To further improve the noise estimation, in the second iteration, the LLRs from the LDPC decoders are fed to the multilevel CNN instead of or in addition to the LLRs from the BCJRs. Moreover, the dotted lines between the upper- and lower-layer BCJRs indicate the possibility of exchanging LLRs between these two detectors to account for residual ILI remaining after the equalizer-separator CNN. These LLRs would be used as soft-decision feedback, which would modify the branch labels used in the BCJR trellis processing without requiring additional trellis states. The turbo-detection system exchanges LLRs between BCJRs, LDPC decoders, and multilevel CNN iteratively to reduce the BER and achieve higher areal density.

C. CNN autoencoder for modulation coding

To ameliorate partial erasures and media noise before the write process, modulation coding is used to avoid writing problematic bit patterns.[48] An example 1D problematic pattern is a short sequence of bits where every two adjacent bits have opposite polarities, e.g., the sequence 01010101. For such sequences, the magnetization of grains within a bit cell is affected by the magnetization of neighboring grains. As the neighboring grains are magnetized in the opposite direction, partial erasures can occur, where the grains are not magnetized or are inadvertently switched to a different polarity. Hence, writing such sequences on the media increases the chances of partial erasures since grains in neighboring bit cells would exert magnetic fields in the opposite direction to the intended polarity. For 2D channels, this problem becomes more severe since grains in the down-track and cross-track directions impact the magnetization of grains in a given bit cell, as illustrated by the pattern on the upper layer in Fig. 13. In Ref. 49, a modulation coding scheme is proposed where select bits in the problematic 2 × 2 patterns are deliberately flipped before the write process. The decoder employs a 2D belief propagation algorithm on the factor graph associated with these patterns to correct the flipped bits. In MLMR, the magnetostatic interactions extend beyond the 2D grid to include grains from different layers, as shown in Fig. 13. Thus, 3D problematic bit patterns should be identified and considered by the modulation coding scheme. Such coding schemes can be difficult to design due to the large number of possible problematic patterns. Also, typical coding schemes assume tractable analytical models that are used to derive the solution. However, such models may not be accurate in practice. Hence, the resulting solutions are inherently sub-optimal.

To overcome these limitations, an autoencoder (AE) CNN can be trained for joint modulation coding and detection/equalization. Such a CNN autoencoder is shown in Fig. 14. The encoder CNN accepts bit sequences to be written and outputs modulation-coded binary sequences. To give the encoder enough degrees of freedom to choose a good set of non-problematic encoded bit sequences, generally, the encoding rate is less than one, i.e., each encoded bit will correspond to less than one original bit. To train the AE, a storage channel approximator is needed for the backpropagation algorithm to compute the channel output given coded input bits. The channel approximator can be realized using analytical models or using a separate CNN that is trained as a model approximator on realistic data. The decoder CNN

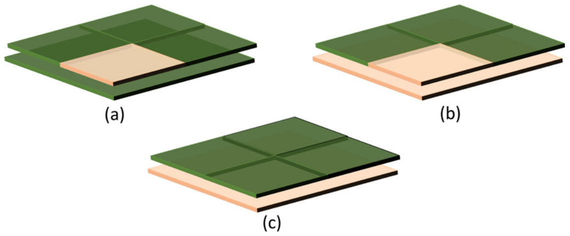

FIG. 13. Example problematic patterns for two-layer recording. Dark and light squares represent bit cells with different polarities. In (a), the magnetization of dark squares on the upper and lower layers undermines the magnetization on the light square. In (b), the light square's magnetization on the lower layer reinforces the light square on the upper layer, but hinders the magnetization of dark squares on the upper layer. In (c), the opposite polarities of bits on different layers can cause partial erasures.

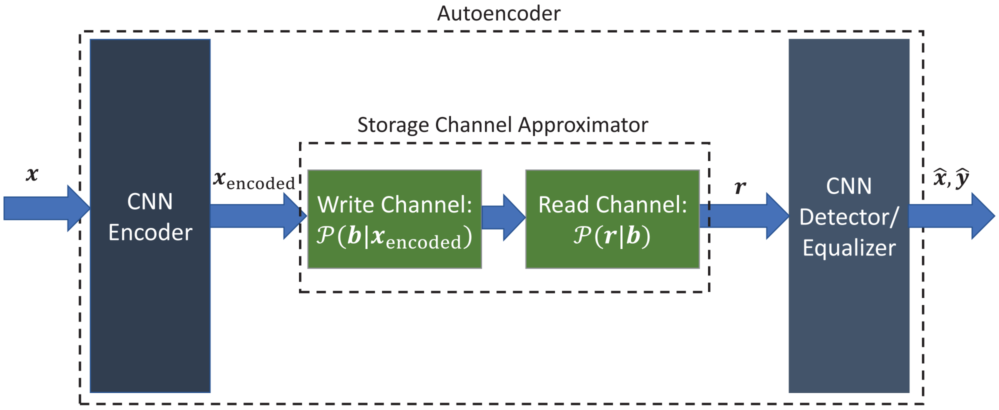
Autoencoder
$\boldsymbol{x}$
CNN Encoder
$\boldsymbol{x}_{\text{encoded}}$
Storage Channel Approximator
Write Channel: $\mathcal{P}(\boldsymbol{b}|\boldsymbol{x}_{\text{encoded}})$
Read Channel: $\mathcal{P}(\boldsymbol{r}|\boldsymbol{b})$
$\boldsymbol{r}$
CNN Detector/ Equalizer
$\hat{\boldsymbol{x}}, \hat{\boldsymbol{y}}$

FIG. 14. CNN autoencoder system. The CNN encoder learns a modulation coding method for encoding the input bit sequence **x**. The encoded binary sequence is written on the hard disk. The storage channel can be approximated by stochastic models for the write and read processes or by a CNN model approximator. The channel output is passed to the CNN decoder for detecting the transmitted bits. The CNN encoder and decoder are trained jointly to minimize the objective function.

can be implemented as a detector (which outputs an estimate $\hat{\mathbf{x}}$ of the input bits $\mathbf{x}$) or an equalizer which outputs a partially equalized signal $\hat{\mathbf{y}}$ for further processing in the receiver. The encoder and decoder CNNs are simultaneously trained to minimize an objective function such as the CE or MSE. Thus, the CNN AE learns the modulation coding and decoding from the data without requiring analytical solutions. An additional possibility at the encoding side of the AE is to integrate the encoder with the write-head signal processing, so that the encoder's output is effectively a write waveform rather than binary bits. Thus, the AE could enable joint optimization of the write waveforms and the receiver processing.

ACKNOWLEDGMENTS

This work was supported by the United States National Science Foundation (NSF) under Grant No. CCF-1817083 and by the Advanced Storage Research Consortium (ASRC).

DATA AVAILABILITY

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

REFERENCES

[1] A. J. Viterbi and J. K. Omura, *Principles of Digital Communication and Coding* (McGraw-Hill, New York, 1979).
[2] R. Wood, M. Williams, A. Kavcic, and J. Miles, IEEE Trans. Magn. **45**, 917 (2009).
[3] R. Galbraith, W. Hanson, B. Lengsfield, T. Oenning, J. Park, and M. Salo, IEEE Trans. Magn. **50**, 67 (2014).
[4] A. R. Krishnan, R. Radhakrishnan, B. Vasic, A. Kavcic, W. Ryan, and M. F. Erden, IEEE Trans. Magn. **45**, 3830 (2009).
[5] C. K. Matcha and S. G. Srinivasa, IEEE Trans. Magn. **51**, 3101215 (2015).
[6] J. Yao, E. Hwang, B. V. K. V. Kumar, and G. Mathew, in *2015 IEEE International Conference on Communications (ICC)* (IEEE, 2015), pp. 418–424.
[7] Y. Wang and B. V. K. V. Kumar, IEEE Trans. Magn. **52**, 3001507 (2016).
[8] Y. Wang and B. V. K. V. Kumar, IEEE Trans. Magn. **52**, 3002011 (2016).
[9] S. Shi and J. R. Barry, in *2018 IEEE International Conference on Communications (ICC)* (IEEE, 2018), pp. 1–6.
[10] Y. LeCun, Y. Bengio, and G. Hinton, Nature **521**, 436 (2015).
[11] K. Luo, S. Wang, K. S. Chan, W. Chen, J. Chen, P. Lu, and W. Cheng, IEEE Trans. Magn. **55**, 6700605 (2019).
[12] Y. Qin and J.-G. Zhu, IEEE Trans. Magn. **56**, 6701108 (2020).
[13] J. Shen, A. Aboutaleb, K. Sivakumar, B. J. Belzer, K. S. Chan, and A. James, IEEE Trans. Magn. **56**, 3100212 (2020).
[14] J. Shen and N. Nangare, *2020 54th Annual Conference on Information Sciences and Systems (CISS)* (IEEE, 2020), pp. 1–6.
[15] M. Nishikawa, Y. Nakamura, Y. Kanai, H. Osawa, and Y. Okamoto, IEEE Trans. Magn. **57**, 3100105 (2021).
[16] A. Sayyafan, B. J. Belzer, K. Sivakumar, J. Shen, K. S. Chan, and A. James, IEEE Trans. Magn. **55**, 6701406 (2019).
[17] A. Sayyafan, A. Aboutaleb, B. J. Belzer, K. Sivakumar, A. Aguilar, C. A. Pinkham, K. S. Chan, and A. James, IEEE Trans. Magn. **57**, 3101113 (2021).
[18] J. Shen, B. J. Belzer, K. Sivakumar, K. S. Chan, and A. James, IEEE Trans. Magn. **57**, 3100905 (2021).
[19] S. K. Nair and J. Moon, IEEE Trans. Neural Networks **8**, 1037 (1997).
[20] A. Aboutaleb, A. Sayyafan, K. Sivakumar, B. Belzer, S. Greaves, K. S. Chan, and R. Wood, IEEE Trans. Magn. **57**, 3101012 (2021).
[21] H. Suto, T. Nagasawa, K. Kudo, K. Mizushima, and R. Sato, Nanotechnol. **25**, 245501 (2014).
[22] S. J. Greaves, K. S. Chan, and Y. Kanai, IEEE Trans. Magn. **55**, 6701509 (2019).
[23] S. J. Greaves, K. S. Chan, and Y. Kanai, IEEE Trans. Magn. **55**, 3001305 (2019).
[24] K. S. Chan, S. Greaves, and S. Rahardja, IEEE Trans. Magn. **55**, 7204905 (2019).
[25] K. S. Chan, R. Wood, and S. Rahardja, IEEE Trans. Magn. **55**, 3002609 (2019).
[26] K. S. Chan, A. Aboutaleb, K. Sivakumar, B. Belzer, R. Wood, and S. Rahardja, IEEE Trans. Magn. **55**, 6701216 (2019).
[27] S. Yoon and E. Hwang, IEEE Trans. Magn. **57**, 3100805 (2021).
[28] Y. Li, Y. Wang, Y. Xu, L. Chen, Y. Wen, and P. Li, IEEE Trans. Magn. **57**, 3100508 (2021).
[29] Z. Yuan and B. Liu, J. Magn. Magn. Mater. **235**, 481 (2001).
[30] M. Albrecht, G. Hu, A. Moser, O. Hellwig, and B. D. Terris, J. Appl. Phys. **97**, 103910 (2005).
[31] N. Amos, J. Butler, B. Lee, M. H. Shachar, B. Hu, Y. Tian, J. Hong, D. Garcia, R. M. Ikkawi, R. C. Haddon *et al.*, PLoS One **7**, e40134 (2012).
[32] S. Khizroev, Y. Hijazi, N. Amos, R. Chomko, and D. Litvinov, J. Appl. Phys. **100**, 063907 (2006).
[33] R. E. Rottmayer, S. Batra, D. Buechel, W. A. Challener, J. Hohlfeld, Y. Kubota, L. Li, B. Lu, C. Mihalcea, K. Mountfield *et al.*, IEEE Trans. Magn. **42**, 2417 (2006).
[34] M. H. Kryder, E. C. Gage, T. W. McDaniel, W. A. Challener, R. E. Rottmayer, G. Ju, Y. Hsia, and M. F. Erden, Proc. IEEE **96**, 1810 (2008).
[35] M. Albrecht, O. Hellwig, G. Hu, and B. D. Terris, "Method for magnetic recording on patterned multilevel perpendicular media using thermal

assistance and fixed write current," U.S. patent 6,865,044 (Hitachi Global Storage Technologies Netherlands B.V., 2005).

[36]H. Yamane, S. J. Greaves, and Y. Tanaka, IEEE Trans. Magn. **57**, 3200706 (2021).

[37]J. G. Zhu, X. Zhu, and Y. Tang, IEEE Trans. Magn. **44**, 125 (2008).

[38]G. Winkler, D. Suess, J. Lee, J. Fidler, M. A. Bashir, J. Dean, A. Goncharov, G. Hrkac, S. Bance, and T. Schrefl, Appl. Phys. Lett. **94**, 232501 (2009).

[39]S. Li, B. Lifshitz, H. N. Bertram, E. E. Fullerton, and V. Lomakin, J. Appl. Phys. **105**, 07B909 (2009).

[40]S. J. Greaves, Y. Kanai, and H. Muraoka, IEEE Trans. Magn. **52**, 3001104 (2016).

[41]J. Hagenauer and P. Hoeher, in *1989 IEEE Global Telecommunications Conference and Exhibition 'Communications Technology for the 1990s and Beyond'* (IEEE, 1989), Vol. 3, pp. 1680–1686.

[42]D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," arXiv:1412.6980 (2014).

[43]N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, J. Mach. Learn. Res. **15**, 1929 (2014).

[44]I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, *Deep Learning* (MIT Press, Cambridge, 2016), Vol. 1.

[45]S. L. Smith, P.-J. Kindermans, C. Ying, and Q. V. Le, "Don't decay the learning rate, increase the batch size," arXiv:1711.00489 (2017).

[46]H. Jin, A. Khandekar, R. McEliece *et al.*, in *Proceeding of the 2nd International Symposium on Turbo Codes and Related Topics* (Citeseer, 2000), pp. 1–8.

[47]L. Bahl, J. Cocke, F. Jelinek, and J. Raviv, IEEE Trans. Inf. Theory **20**, 284 (1974).

[48]B. H. Marcus, P. H. Siegel, and J. K. Wolf, IEEE J. Sel. Areas Commun. **10**, 5 (1992).

[49]M. Bahrami and B. Vasić, IEEE Trans. Commun. **68**, 752 (2020).