Korean Chem. Eng. Res., 62(1), 118-124 (2024)
https://doi.org/10.9713/kcer.2024.62.1.118
PISSN 0304-128X, EISSN 2233-9558

잡음이 포함된 측정 자료에 대한 신경망의 DNA 용액 조성비 예측

강경희*,‡ · 김민지*,‡ · 이효민†

제주대학교 화학공학과
63243 제주시 제주대학로 102
(2023년 12월 4일 접수, 2023년 12월 21일 수정본 접수, 2023년 12월 21일 채택)

Prediction of Composition Ratio of DNA Solution from Measurement Data with White Noise Using Neural Network

Gyeonghee Kang*,‡, Minji Kim*,‡ and Hyomin Lee†

Department of Chemical Engineering, Jeju National University, 102 Jejudaehak-ro, Jeju-si, 63243, Korea
(Received 4 December 2023; Received in revised from 21 December 2023; Accepted 21 December 2023)

요 약

신경망은 심전도 신호, 망막 영상, 지진파 등 잡음이 포함된 자료의 전처리 작업에 활용되고 있다. 그러나, 잡음의 전처리는 전산시간 증가, 원본 신호의 왜곡등의 문제점을 내포하고 있다. 본 연구에서는 잡음의 전처리 없이 측정 자료를 분석할 수 있는 신경망 구조를 연구하였다. 신경망의 학습 자료로써 잡음이 포함된 DNA 용액의 동역학적 거동을 선정하여, 해당 자료로부터 DNA 용액의 조성비를 예측하고자 하였다. DNA의 동역학 자료에 인위적으로 백색 잡음을 추가하여, 신경망의 예측에 대한 잡음의 영향을 알아보았다. 결과적으로, 잡음의 전처리 없이 $O(1)$의 신호 대 잡음비 자료로부터 $O(0.01)$의 오차로 용액의 조성비를 예측할 수 있었다. 이러한 연구 결과는 측정 잡음에 민감하게 영향 받을 수 있는 극미량의 유전병 또는 암세포와 관련된 DNA를 분석을 위한 핵심 인공지능 기술로 활용할 수 있다.

Abstract – A neural network is utilized for preprocessing of de-noizing in electrocardiogram signals, retinal images, seismic waves, etc. However, the de-noizing process could provoke increase of computational time and distortion of the original signals. In this study, we investigated a neural network architecture to analyze measurement data without additional de-noizing process. From the dynamical behaviors of DNA in aqueous solution, our neural network model aimed to predict the mole fraction of each DNA in the solution. By adding white noise to the dynamics data of DNA artificially, we investigated the effect of the noise to neural network's predictions. As a result, our model was able to predict the DNA mole fraction with an error of $O(0.01)$ when signal-to-noise ratio was $O(1)$. This work can be applied as a efficient artificial intelligence methodology for analyzing DNA related to genetic disease or cancer cells which would be sensitive to background measuring noise.

Key words: Neural network, Prediction mole fraction of DNA, White noise, Signal-to-noise-ratio

1. 서 론

최근 신경망을 활용한 공학적 예측 작업 수행이 다양한 분야에서 사용되고 있다. 유체역학 분야에서는 수식-기반 모델링(equation-based modeling)의 고전적인 방식을 대체하는 데이터-기반 모델링(data-driven modeling)을 통해 항력 감소 문제(drag reduction), 유동 특성 추출(flow feature, extraction), 유동장 분석, 유동 최적화 등의 문제에 신경망 기법을 적용하고 있다[1]. 열역학 분야에서는 상평형(phase eqiulibria)의 분석을 신경망으로 처리한 연구가 보고되었으며[2-5], 전기전자공학 분야에서는 단층촬영(tomography)의 영상처리에 신경망을 활용하였다[6].

다양한 공학 영역 중, 잡음이 섞인 데이터를 다루는 연구에서는 일반적으로 잡음의 전처리 작업이 선행된다. 예를 들어, 심전도 신호의 전력선 잡음 제거[7], 오토인코더 구조를 활용한 망막 영상의 잡음 개선[8], 무선 신호처리의 잡음 제거[9], 지진파 잡음 제거[10] 등이 있다. 이러한 작업은 DNA와 단백질 등을 분석하는 미세 유체 기술에도 잡음 제거의 선행이 필요할 수 있다.

미세 유체 기술 중 DNA 미세배열(microarray)은 동시에 수천 개의 유전자 발현 상황을 탐색케 한다. 미세배열 장치에 부착된 상보적 DNA(cDNA, complementary DNA)의 형광 신호를 이미지 형태

†To whom correspondence should be addressed.
E-mail: fluid@jejunu.ac.kr
‡Those authors contributed equally.

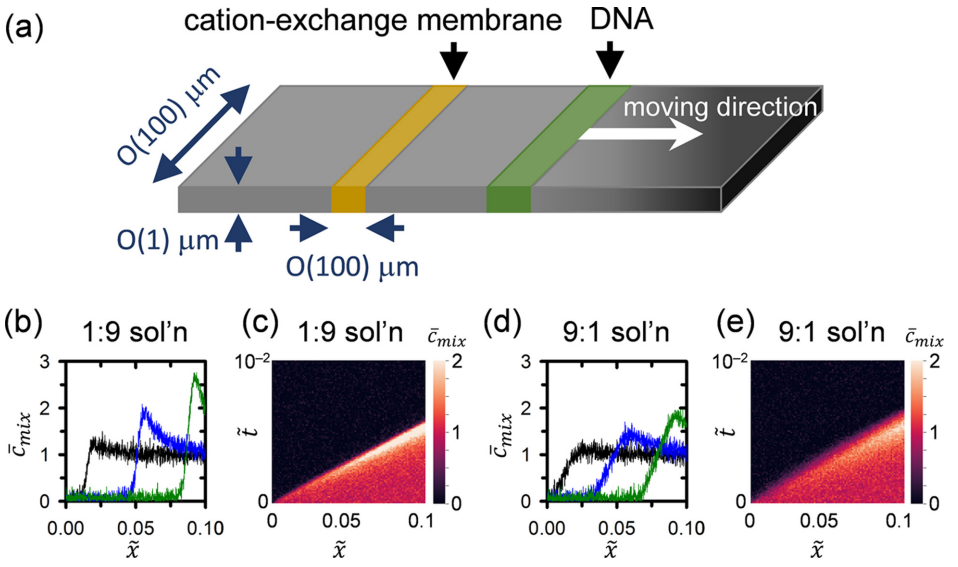

Fig. 1. (a) Schematic of microfluidic platform for DNA manipulation. (b) Concentration profiles for DNA solution of $c_{10\,bp}^{bulk}:c_{1000\,bp}^{bulk}$=1:9. (c) Spatiotemporal map for DNA solution of . (d) Concentration profiles for DNA solution of $c_{10\,bp}^{bulk}:c_{1000\,bp}^{bulk}$=9:1. (e) Spatiotemporal map for DNA solution of $c_{10\,bp}^{bulk}:c_{1000\,bp}^{bulk}$ = 9:1. In (b) and (d), each colored solid line corresponded with $\tilde{t}$ = 0.001 (black), 0.0025 (blue), 0.005 (green), respectively.

의 자료로 표현하여, cDNA의 형태와 양을 분석하게 된다. 이러한 실험은 이미지 처리 과정에서 의도하지 않은 잡음이 포함될 수 있다. 따라서 분석 과정 중의 잡음을 인식하고 제거하기 위한 전처리가 필요하다[11,12]. 잡음 전처리를 위한 일반적인 방법론은 평균 필터(mean filter)[13], 중앙값 필터(median filter)[14], 가우스 필터(Gaussian filter)[15]와 같은 필터링 알고리즘을 이용한다.

기존의 잡음 전처리 방법론은 전산 시간(computation time)의 증가 혹은 원본 신호 왜곡의 문제점을 내포한다. 이러한 문제를 극복하기 위해 본 연구에서는, 잡음 전처리 과정을 생략하여 잡음이 포함된 동역학 자료를 바로 분석할 수 있는 신경망을 개발하고자 하였다. 동역학 자료는 수치 모델(numerical model)을 통해 얻었으며, 수치 해석 영역(numerical domain)은 Fig. 1(a)와 같은 $O(1)$ cm × $O(100)$ mm × $O(1)$ mm 크기의 미세 유로 장치를 고려하였다. 해당 영역에서의 분석 물질 거동은 본 연구진에 의해 분석하여 보고되었다[16]. 미세 유로 장치에 $O(100)$ mm의 양이온 교환 막을 배치하고 구동 전류를 가하면, 양이온 교환 막은 막/유체 계면에 이온 농도 분극(ion concentration polarization)을 유발하여 농도장, 유동장, 전기장이 국소적으로 변화한다[16-20]. 이에 따라, 10 bp와 1,000 bp의 DNA 용액은 조성에 따라 서로 다른 거동을 보인다. 수치 모델을 통해 계산되어진 DNA 동역학 자료는, Fig. 1(b)-(e)와 같이, 공간시간 지도로 시각화하여 신경망의 입력 자료로 활용하여 용액의 DNA 조성을 추정하는 신경망을 개발하였다. 결과적으로, 잡음의 전처리 과정이 없어 잡음이 포함된 자료가 분석 대상이 되었음에도, DNA 조성의 예측 오차는 $O(0.01)$ 몰 분율로 정확한 예측을 할 수 있었다. 본 연구의 신경망을 활용한다면 잡음이 포함된 측정 자료를 분석함에 있어 잡음의 전처리 과정을 생략할 수 있으며, 측정 잡음에 민감하게 영향 받을 수 있는 극미량의 유전병 또는 암세포와 관련된 DNA를 분석을 위한 핵심 인공지능 기술로 활용할 수 있다.

2. 기계 학습 방법론

2-1. 심층 신경망

Fig. 2(a)는 본 연구에서 사용된 심층 신경망 구조를 도식화한 것이다. 합성곱 신경망(convolutional neural network)과 완전-연결 신경망(fully-connected neural network)을 활용하여, DNA 혼합 용액의 동역학적 공간시간 지도를 분석하였다. 동역학적 공간시간 지도는 입력층을 통해 신경망으로 입력된다. 합성곱 신경망은 공간 차원을 줄이고 중요한 특징을 추출하도록 하기 위해, 합성곱 층(convolution layer)과 최대-풀링 층(max-pooling layer)을 교차로 구성하였다. 이 때 합성곱 층은 3 × 3 크기의 커널(kernel)을 사용하였으며, 세 개의 합성곱 층은 각각 8, 16, 32개의 필터를 설정하였다. 최대-풀링 층은 2 × 2 크기의 풀링 크기를 사용하였다. 그 다음 완전 연결 신경망은 5개의 은닉층(hidden layer)으로 구성하였다. 각 층의 뉴런은 각각 480, 240, 120, 60, 30개로 설정하였다. 합성곱 층과 은닉층의 활성화 함수(activation function)는 ReLU (Rectified Linear Unit)함수를 도입하였다. 출력층은 1개의 뉴런을 두어 DNA 용액에 포함된 10 bp DNA의 예측 몰분율을 출력하도록 하였다. 신경망의 목표값(target)이 0.5 이하인 경우, 예측 정확도가 떨어지는 문제를 해결하기 위하여 출력층의 활성화 함수는 LeakyReLU (Leaky Rectified Linear Unit)함수를 사용하였다. 이러한 구조로 구성된 신경망은 adam (adaptive moment estimation) 최적화 알고리즘을 활용하여 신경망을 학습시켰다. Fig. 2(b)와 같이, 본 연구의 신경망은 과적합 없이 적합한 훈련 성능 보이는 것을 확인하였다. 훈련 횟수(epochs)가 350회 이상부터 평균 제곱근 오차(Root-Mean-Squared-Error, RMSE)가 일정 값으로 수렴하므로, 훈련 횟수를 500회로 고정하였다.

신경망의 기계 학습은 훈련 횟수와 훈련 자료의 개수가 많아질수록 학습 성능은 더 향상된다. 최적의 훈련 자료의 개수를 결정하기 위하여, Fig. 2(c)와 같이, 훈련 자료의 개수에 따른 신경망의 학습 성능을 분석하였다. 훈련 자료의 개수가 200개 이상인 경우부터는

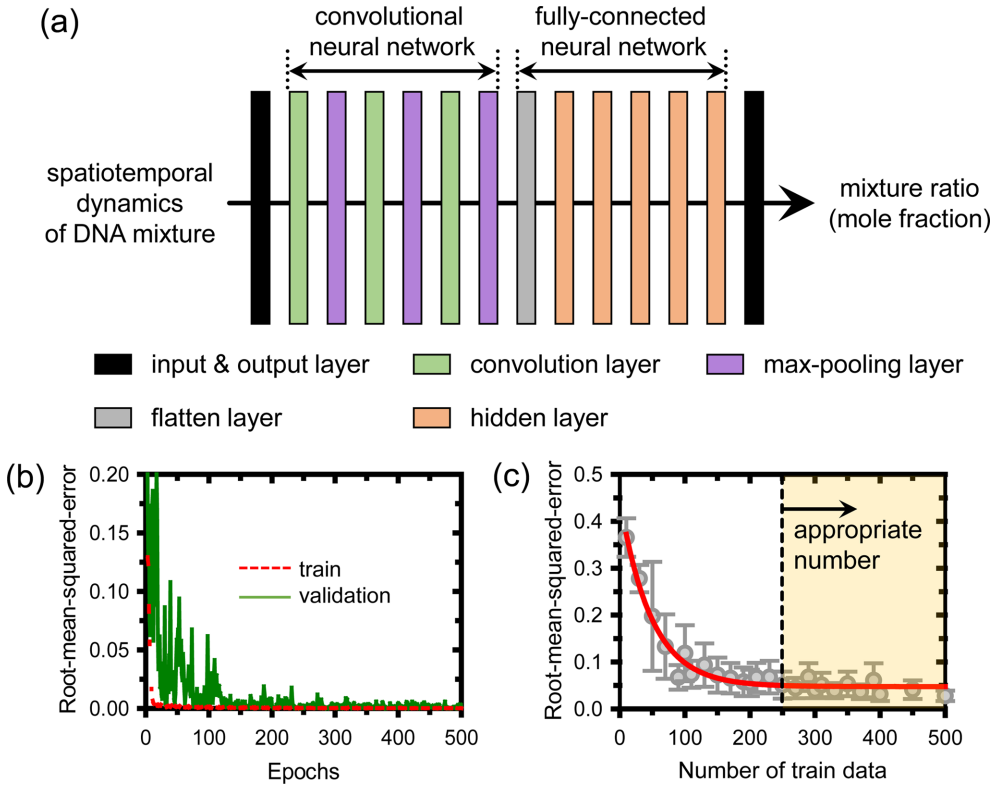

Fig. 2. (a) Configuration of deep neural network structures. (b) Root-mean-squared-error for train and validation data as a function of number of epochs. Red dashed line and green solid line correspond with train and validation data, respectively. (c) Root-mean-squared-error affected by number of train data. Symbols and error bars represent average and standard deviation at the given number of train data, respectively. Red solid line is regression for data point of symbols. The shaded region shows appropriate number of train data.

RMSE가 일정 값으로 수렴하며, 이를 고려하여 적정 훈련 자료의 개수를 250개로 결정하였다. 따라서 본 연구에서는 총 500번의 훈련 횟수와 250개의 훈련 자료를 사용하여 전체 전산 실험을 수행하였다.

2-2. DNA 동역학

심층 신경망의 학습 대상인 미세유로에서의 DNA 동역학은 유한요소법(finite element method)기반의 수치 해석을 통해 구성하였다. Fig. 1(a)와 같이, 왼쪽 끝은 양이온-교환막(cation-exchange membrane)으로 막혀있고, 오른쪽 끝은 벌크 저수조(bulk reservoir)에 연결된 막다른 미세유로를 수치 해석 영역으로 설정하였다. 이러한 미세유체계는 분석 물질의 농축 및 분리[21], 전기동역학적 특성 분석[22] 등에 활용된다.

막다른 미세유로에서 이온과 분석 물질의 거동은 면-평균 모델(area-averaging model)기반의 수치적 조직화(numerical formulation)[16,23]를 이용하여 계산하였다. Fig. 1(a)의 영역에 대하여, 다음의 무차원 지배 방정식이 성립한다.

$$\tilde{c} \equiv \tilde{c}_- = \tilde{c}_+ + 2\tilde{\rho}_s \tag{1}$$

$$\frac{\partial \tilde{c}}{\partial \tilde{t}} = -\frac{\partial}{\partial \tilde{x}}\left(-\frac{\partial \tilde{c}}{\partial \tilde{x}} + \tilde{c}\frac{\partial \tilde{\phi}}{\partial \tilde{x}}\right) \tag{2}$$

$$\frac{\partial}{\partial \tilde{x}}\left[-(\tilde{c}-\tilde{\rho}_s)\frac{\partial \tilde{\phi}}{\partial \tilde{x}}\right] = 0 \tag{3}$$

여기에서 틸데(tilde, '~')기호와 함께 표기된 변수는 무차원임을 의미하며, 무차원화를 위한 크기 척도는 Table 1에 표시하였다. 식 (1)은

Table 1. Used characteristic scale for non-dimensional variables

Quantity	Related variable	Characteristic scale	Description
Time	t	$\tau_D \equiv \frac{L^2}{D}$	Diffusion time scale
Length	x	L	Distance from CEM to bulk
Electric potential	ϕ	$V_T \equiv \frac{RT}{F}$	Thermal voltage
Concentration	c	c_0	Bulk concentration
Charge	ρ_s	$\rho_0 \equiv 2Fc_0$	Bulk-scaled charge concentration
Current density	i	$i_0 \equiv \frac{2FDc_0}{L}$	Diffusion-limited current density

국소 전기중성성(local electroneutrality)이고, 식 (2)는 배경 전해질(background electrolyte)의 질량 보존 법칙, 식 (3)은 이온 전류의 보존 법칙이다. 위 식에서 $\tilde{\rho}_s$는 미세유로의 무차원 표면 전하 밀도(surface charge density)로써, 다음과 같이 정의된다.

$$\tilde{\rho}_s \equiv \frac{a_p q_s}{2AFc_0} \tag{4}$$

분석 물질인 DNA의 동역학은 다음의 물질 전달 방정식으로 서술된다.

$$\frac{\partial \tilde{c}_i}{\partial \tilde{t}} = -\frac{\partial}{\partial \tilde{x}}\left(-\tilde{D}_i\frac{\partial \tilde{c}_i}{\partial \tilde{x}} - \tilde{\mu}_i \bar{c}_i \frac{\partial \tilde{\phi}}{\partial \tilde{x}}\right) \tag{5}$$

여기에서 아래 첨자 i는 DNA의 길이를 의미하여, 10 bp와 1,000 bp 두 종류의 DNA 동역학을 계산하였다. 또한, $\bar{c}_i \equiv c_i/(c_{10bp}^{bulk} + c_{1000bp}^{bulk})$, $\tilde{D}_i \equiv D_i/D$, $\tilde{\mu}_i \equiv \mu_i/\mu_0$ 이다. 10 bp DNA의 확산계수와 전기영동 이동도는 각각 $9.667\times10^{-11}\ m^2s^{-1}$, $3.328\times10^{-8}\ m^2V^{-1}s^{-1}$를 사용하였고, 1,000 bp DNA는 $6.97111907\times10^{-12} m^2s^{-1}$, $3.75909848\times10^{-8}\ m^2V^{-1}s^{-1}$의 확산계수와 전기영동 이동도를 사용하였다[24,25]. 식 (1)-(5)의 수치적 조직화의 자세한 내용은 참고문헌[16,23]에 정리되어 있다.

해석 영역에 대한 경계조건으로써, 벌크 저수조 경계에서 일정 전류 밀도, 벌크 이온 농도, 벌크 DNA 농도를 설정하였다.

$$(1-\tilde{\rho}_s)\frac{\partial \tilde{\phi}}{\partial \tilde{x}} = \tilde{i}_{app} \tag{6}$$

$$\tilde{c} = 1 \tag{7}$$

$$\bar{c}_i = \bar{c}_i^{bulk} \tag{8}$$

양이온-교환막 표면은 전기적 기저(electrical ground), 이상적인 양이온 선택성, DNA에 대한 불투과성 조건을 선택하였다.

$$\tilde{\phi} = 0 \tag{9}$$

$$-\frac{\partial \tilde{c}}{\partial \tilde{x}} + \tilde{c}\frac{\partial \tilde{\phi}}{\partial \tilde{x}} = 0 \tag{10}$$

$$-\tilde{D}_i\frac{\partial \tilde{c}_i}{\partial \tilde{x}} - \tilde{\mu}_i \bar{c}_i \frac{\partial \tilde{\phi}}{\partial \tilde{x}} = 0 \tag{11}$$

면-평균 모델이므로, 미세유로의 벽면에 대한 불투과성 조건은 지배방정식에 포함되었다.

식 (1)-(11)은 COMSOL Multiphysics 5.6을 이용하여 완전-결합 분석(fully-coupled analysis)으로 풀었다. 수치 해석으로 구성한 학습 자료에 Gauss 잡음을 인위적으로 섞어 실제 실험의 광학 측정 자료와 유사하게 되도록 하였다. Fig. 1(b)는 $c_{10bpi}^{-bulk} + c_{1000bp}^{-bulk}$ = 1:9 인 혼합 용액을 대상으로 계산된 시간에 따른 DNA의 총 농도 분포(total concentration distribution) 결과이다. 신호대잡음비(signal to noise ratio, SNR)를 10으로 설정한 결과, 농도 분포에 백색 잡음이 합쳐져 나타나는 것을 알 수 있다. 이러한 DNA의 동역학은 신경망으로 입력될 때, Fig. 1(c)의 공간시간 지도(spatiotemporal map)로 변환되도록 하였다. Fig. 1(d), (e)는 $c_{10bpi}^{-bulk} + c_{1000bp}^{-bulk}$ = 9:1인 혼합 용액의 농도 분포와 공간시간 지도를 나타내었으며, 1:9 용액과 비교하여 동역학의 양상이 달라져 있음을 알 수 있다. 따라서, 본 연구의 신경망의 학습 목표는 공간시간 지도로 입력된 동역학의 변화 양상을 분석하여 혼합 용액의 물리화학적 정보를 결정하는 것으로 설정하였다.

3. 결과 및 고찰

3-1. 잡음이 포함된 동역학계 분석 및 성능 포화 지점

동역학계의 공간시간 지도(spatiotemporal map)는 공간과 시간 정보를 결합하여 시각적으로 나타낸 단일 도표를 의미한다. 공간시간 지도 상에서는, 데이터의 해석과 분석이 용이하며, 동역학의 주된 변화를 쉽게 파악할 수 있다. 따라서, 해당 지도는 데이터의 복잡성을 단순화하고 숨겨진 상관 관계나 경향성을 찾아내는데 이용된다. 또한, 공간시간 지도는 화학 물질의 정성적 및 정량적 특성화 및 분석, 난류 흐름 분석, 담수화 공정 해석에 유용한 것으로 밝혀져 있다[26-28]. 본 연구에서는 시료의 조성별로 달라지는 동역학을 분석하고자, 측정 가능한 거동의 공간시간 지도를 활용하였다. 본 연구의 공간시간 지도에는 실험에서 검출될 수 있는 잡음을 포함하였다. 해당 잡음의 전처리 없이 신경망을 이용하여 지도를 분석하여, 측정 가능한 동역학 거동으로부터 시료의 조성을 예측하고자 하였다.

분석하기 위한 공간시간 지도는 10 bp와 1,000 bp가 DNA 혼합 용액의 무차원 구동 전류 20에서의 동역학으로부터 작성하였다. Fig. 3(a)-3(c)의 지도는 신호-대-잡음비(Signal-to-Noise Ratio, SNR)에 따라 변형된 공간시간 지도를 나타내며, 지도의 색상 정보는 DNA 혼합물의 총 농도를 의미한다. Fig. 3(a)와 같이 측정 신호 대비 잡음이 매우 큰 경우(SNR<<1), 신경망을 통한 분석은 신호와 잡음을 구분하지 못하여 예측의 정확도가 떨어진다. SNR<<1일 때, 잡음이 신호에 비해 우세하게 나타나므로, 신호를 구분할 수 없다. 반면, Fig. 3(b)와 3(c)와 같이, SNR ≥ 1이면 DNA의 농축 및 이동을 의미하는 신호가 잡음으로부터 구분이 가능하며, 신경망을 통한 지도의 분석을 기대해볼 수 있다. 전형적인 센서의 검출 한계(limit of detection)의 SNR은 $O(10^{-2})$이므로[29], 본 연구에서는 $10^{-2} \le$ SNR $\le 10^2$의 동역학 자료를 통해 신경망의 예측 성능의 변화 여부를 알아보았다.

예측 성능을 정량화하고자, 신경망의 목표 값(target value)과 예측 값(predicted value) 간의 평균 제곱근 오차(Root-Mean-Squared-Error, RMSE)를 이용하였다.

$$\mathrm{RMSE} = \sqrt{\sum_i \frac{(y_i - \hat{y}_i)^2}{N}} \tag{12}$$

작은 RMSE 값은 목표 값과 예측 값 간의 차이가 작음을 의미하여, 인공지능의 예측 성능이 높음을 나타낸다. 예를 들어, DNA의 몰 분율을 예측하는 과업의 경우, $O(0.01)$의 RMSE는 예측 몰 분율이 적어도 2개의 유효 숫자를 가질 수 있음을 의미한다. 본 연구의 신경망 분석은 Fig. 3(d)에 나타난대로, $O(0.01)$의 RMSE로 정량화된 성능을 얻었다.

Fig. 3(d)에서, SNR이 특정 값 이상이 되면 오차가 더 이상 줄어들지 않고 성능이 포화되는 영역이 존재한다. 성능 포화가 시작되는 신호-대-잡음비를 SNR_{lim}로 정의하였고, RMSE 곡선의 감소하는 부분의 접선과 수렴하는 부분의 접선이 만나는 지점(*cf.* 파란 점선)으로 정하였다. 구동 전류가 20일 때, SNR_{lim}은 1.775이다. 따라서, SNR > SNR_{lim}이면 포화된 성능으로 예측을 수행할 수 있다. 반면, SNR << SNR_{lim} 이면, 잡음과 신호의 구분이 불가능하여 예측의 오차가 증가하게 된다.

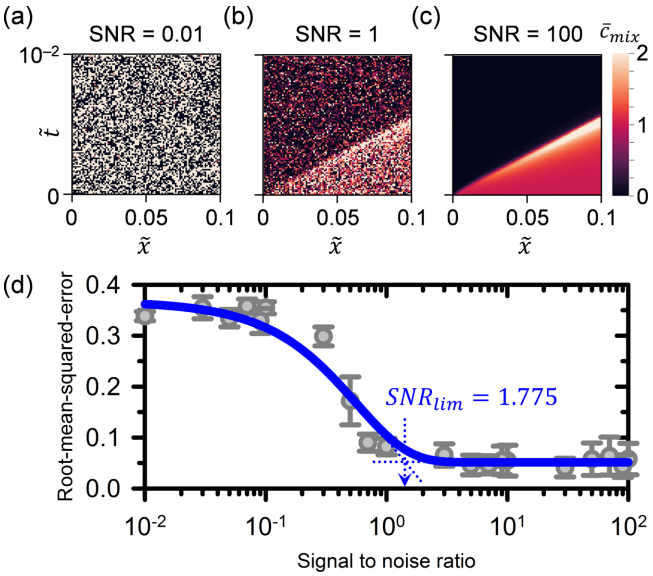

Fig. 3. The spatiotemporal map of the DNA dynamics when dimensionless applied current density was 20 at (a) SNR = 0.01, (b) 1, (c) 100. The represented colors in the map were the total concentration of the DNA solution. (d) Root-mean-squared-error with respect to signal-to-noise ratio for validation data. SNR_{lim} was 1.775. When the signal-to-noise-ratio exceeds the SNR_{lim}, the performance of neural network is independent of the white noise.

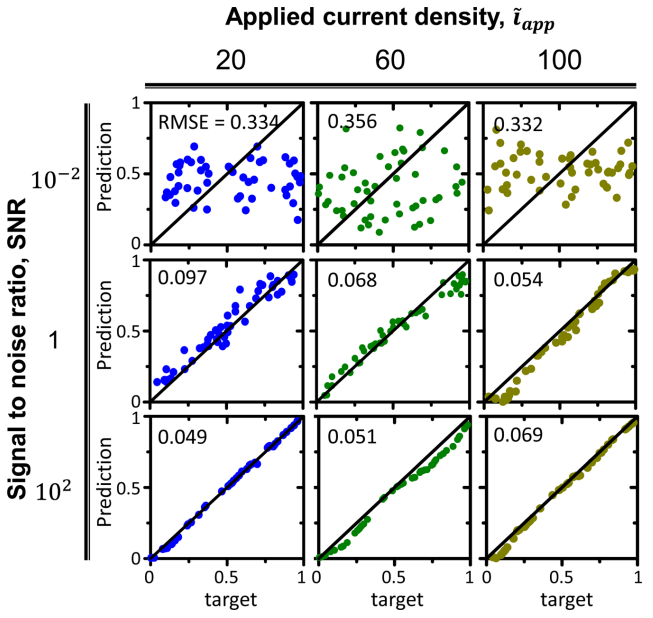

Fig. 4. Confusion matrices with respect to SNR and $\tilde{\imath}_{app}$.

3-2. 분석 조건에 따른 예측 정확

주어진 동역학계에서 구동 전류 밀도(applied current density)의 변화에 따라 DNA 용액의 조성 예측이 영향을 받는다. Fig. 4는 구동 전류 밀도와 SNR에 따른 모델의 예측 정확도를 오차 행렬(confusion matrix)로 가시화한 결과를 보여준다. SNR 값이 매우 낮을 때, 즉 SNR = 10^{-2}인 경우에는 신호 대비 잡음이 매우 크기 때문에, 구동 전류 밀도를 높여도 인공지능 모델의 예측이 부정확하다. 부정확의 정도는 그래프의 검은 실선(인공지능의 목표 값 가시화)을 중심으로 멀리 흩어진 분포를 통해 알 수 있다. 검은 실선과 인공지능의 예측 값의 차이를 평균 제곱근 오차(RMSE)로 계산하여 각 그래프에 나타내었다. 반면, SNR 값이 1일 때는 SNR = 10^{-2}인 경우와는 다른 특징을 보인다. SNR = 10^{-2}인 경우에 비해, 인공지능 모델의 예측은 목표 값 근처에서 일어난다. 더불어, 구동 전류가 증가함에 따라 RMSE가 작아지면서 예측 성능이 점차 향상된다. 따라서, 구동 전류가 높아질수록 예측에 대한 잡음의 영향을 감소시킬 수 있다. 또한, SNR = 10^{-2}인 경우, 모든 구동 전류 값 범위에서 모델은 우수한 예측을 수행한다. 그러나 구동 전류가 커질수록 RMSE가 커지는 것을 확인할 수 있는데, 이는 모델이 훈련 데이터를 과하게 학습하여 과적합이 발생한 결과이다. 이러한 결과를 통해 성능 포화 지점이 구동 전류에 영향을 받음을 확인할 수 있다.

3-3. 분석 조건에 따른 성능 포화 지점의 변화

Fig. 5(a)는 구동 전류 밀도별 SNR에 따른 RMSE를 나타내어, 구동 전류에 따른 성능 포화 지점인 SNR_{lim}를 나타낸 것이다. 동역학계의 신호대잡음비가 SNR_{lim}보다 크다면, 인공지능의 예측은 잡음에 영향을 받지 않게된다. 구동 전류 밀도가 20에서 100으로 높아질수록 성능 포화 지점인 SNR_{lim}은 1.775에서 0.841로 감소한다. 다시 말해, 높은 구동 전류 밀도의 조건에서는 더 큰 잡음의 환경을 분석할 수 있다. SNR → 0일 때, RMSE가 수렴하는 것은 잡음이 너무 크기 때문에 모델이 유용한 정보나 패턴을 학습하지 못하고 무작위로 예측하기 때문이다. 무작위 예측인 경우의 최소 오차는 0.5 이므로, $\lim_{SNR\to 0}$ RMSE = 0.5의 거동이 모든 그래프에서 나타난다.

Fig. 5(b)는 구동 전류 밀도에 따른 성능 포화 지점을 시각적으로 나타낸 것이다. 이를 통해 구동 전류 밀도가 높아질수록 전체적으로 분석 성능이 향상되는 경향이 확인 가능하다. 그러나 성능은 특정 지점에서 포화되며, 이 때의 SNR_{lim}는 0.661이다. 이는 구동 전류를 높여도 성능 향상이 더 이상 일어나지 않는 지점을 나타낸다. 이러한 결과는 구동 전류를 변경하여 성능을 향상시키는 방법에는 한계가 있음을 나타낸다. 따라서, 모델의 성능을 향상시키기 위해서는 신경망의 구조 개선, 학습 알고리즘 교체 등의 방법을 고려해야한다.

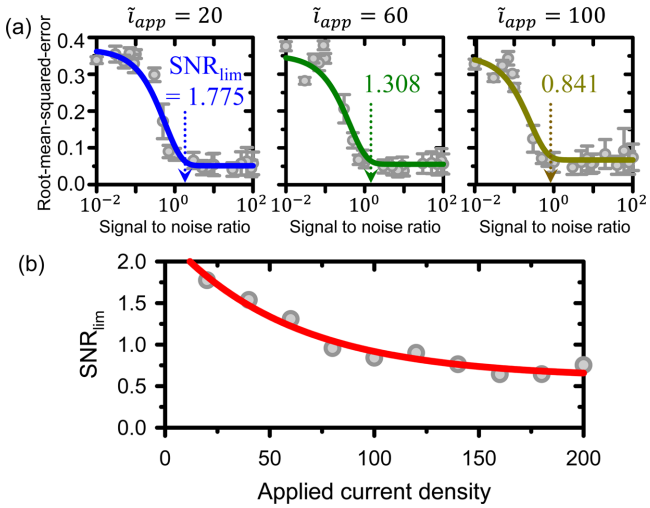

Fig. 5. (a) Root-mean-squared-error with respect to signal-to-noise ratio for each applied current density. (b) SNR_{lim} as a function of applied current density.

4. 결 론

본 연구에서는 신경망을 활용해 잡음이 포함된 DNA 용액의 측정자료를 분석하여, 10 bp와 1,000 bp로 구성된 DNA 혼합 용액의 조성비를 예측하는 작업을 수행하였다. 이를 위해, 신호-대-잡음비(SNR)에 따라 변형되는 공간시간 지도를 신경망의 학습 자료로 활용하였으며, 잡음의 전처리 과정 없이 신경망으로 신호와 잡음을 모두 분석하였다. 그 결과, 일정 수준의 신호-대-잡음비까지는 $O(0.01)$의 정확도로 DNA의 몰분율을 예측할 수 있었다. 신경망으로 분석이 가능한 신호-대-잡음비의 한계를 SNR_{lim}으로 정의하였으며, SNR_{lim}은 구동 전류 밀도의 증가에 따라 감소하는 경향을 확인하였다. 다시 말해, 동역학계의 구동 전류 밀도가 커진다면, 신경망을 통해 측정 신호와 더 큰 잡음을 전처리없이 분석 가능함을 보였다. 본 연구의 신경망을 활용한다면, 잡음이 포함된 측정 자료를 분석함에 있어 잡음의 전처리 과정을 생략할 수 있다. 이러한 연구 결과는 측정 잡음에 민감하게 영향 받을 수 있는 극미량의 유전병 또는 암세포와 관련된 DNA를 분석을 위한 핵심 인공지능 기술로 활용할 수 있다.

감 사

이 논문은 2022학년도 제주대학교 교원성과지원사업에 의하여 연구되었습니다.

사용 기호

A : Cross-sectional area of microchannel
a_p : Wetted perimeter of microchannel
c : Ionic strength of background electrolyte [mol m^{-3}]
c_0 : Bulk concentration of background electrolyte [mol m^{-3}]
c_i : Concentration of background electrolyte of *i*-th ionic species [mol m^{-3}]
c_{10bp}^{bulk} : Bulk concentration of 10 bp DNA
c_{1000bp}^{bulk} : Bulk concentration of 1,000 bp DNA
$\tilde{c}$: Dimensionless ionic strength of background electrolyte
$\tilde{c}_+$: Dimensionless concentration of background cation
$\tilde{c}_-$: Dimensionless concentration of background anion
$\bar{c}_i$: Dimensionless concentration of *i*-th DNA
c_i^{-bulk} : Dimensionless bulk concentration of *i*-th DNA
D : Characteristic diffusivity [m^2 s^{-1}]
D_i : Diffusivity of *i*-th ionic species [m^2 s^{-1}]
$\tilde{D}_i$: Dimensionless diffusivity of *i*-th ionic species
F : Faraday constant [C mol^{-1}]
ϕ : Electric potential [V]
$\tilde{\phi}$: Dimensionless electric potential
i : Current density [A m^{-2}]
i_0 : Applied current density [A m^{-2}]
$\tilde{\imath}_{app}$: Dimensionless applied current density
L : Distance from CEM to bulk
μ_0 : Characteristic electrophoretic mobility [m^2 V^{-1} s^{-1}]
μ_i : Electrophoretic mobility of *i*-th species [m^2 V^{-1} s^{-1}]
$\tilde{\mu}_i$: Dimensionless electrophoretic mobility of *i*-th species
N : Number of data
q_s : Surface charge density of microchannel wall [C m^{-2}]
R : Gas constant [J mol^{-1} K^{-1}]
ρ_0 : Bulk-scaled charge concentration [C m^{-3}]
ρ_s : Surface charge concentration of microchannel wall [C m^{-3}]
$\tilde{\rho}_s$: Dimensionless surface charge concentration
T : Absolute temperature [K]
t : Time [s]
$\tilde{t}$: Dimensionless time
τ_D : Diffusion time scale [s]
V_T : Thermal voltage [V]
x : Spatial coordinate in *x*-direction [m]
$\tilde{x}$: Dimensionless spatial coordinate
y_i : Predicted mole fraction of 10 bp DNA by neural network
$\hat{y}_i$: Target mole fraction of 10 bp DNA for neural network

Reference

1. Brunton, S. L., Noack, B. R. and Koumoutsakos, P., "Machine Learning for Fluid Mechanics," *Annu. Rev. Fluid Mech.*, **52**, 477-508(2020).
2. Ghavipour, M., Ghavipour, M., Chitsazan, M., Najibi, S. H. and Ghidary, S. S., "Experimental Study of Natural Gas Hydrates and a Novel Use of Neural Network to Predict Hydrate Formation Conditions," *Chemical Engineering Research and Design*, **91**, 264-273(2013).
3. Landgrebe, M. K. B. and Nkazi, D., "Toward a Robust, Universal Predictor of Gas Hydrate Equilibria by Means of a Deep Learning Regression," *ACS Omega*, **4**, 22399-22417(2019).
4. Poort, J. P., Ramdin, M., van Kranendonk, J. and Vlugt, T. J. H., "Solving Vapor-liquid Flash Problems Using Artificial Neural Networks," *Fluid Phase Equilibria*, **490**, 39-47(2019).
5. Sun, G. *et al.*, "Vapor-liquid Phase Equilibria Behavior Prediction of Binary Mixtures Using Machine Learning," *Chemical Engineering Science*, **282**, 119358(2023).
6. Hamilton, S. J. and Hauptmann, A., "Deep D-Bar: Real-Time Electrical Impedance Tomography Imaging With Deep Neural Networks," *IEEE Transactions on Medical Imaging*, **37**, 2367-2377(2018).
7. Kwon, O., Leejieun, Hwan, K. J., Seongjun, L. and Yoo, S. K., "Design of Deep De-nosing Network for Power Line Artifact in Electrocardiogram," *Journal of Korea Multimedia Society*, **23**, 402-411(2020).
8. Badar, M., Haris, M. and Fatima, A., "Application of Deep Learning for Retinal Image Analysis: A Review," *Computer Science Review*, **35**, 100203(2020).
9. Zhang, J. *et al.* in *2019 IEEE 89th Vehicular Technology Conference (VTC2019-Spring).* 1-5.
10. Li, Y. and Ma, Z., "Deep Learning-based Noise Reduction for Seismic Data," *Journal of Physics: Conference Series*, **1861**, 012011(2021).
11. Hong, H., Hong, Q., Liu, J., Tong, W. and Shi, L., "Estimating Relative Noise to Signal in DNA Microarray Data," *International*

Journal of Bioinformatics Research and Applications, **9**, 433-448(2013).
12. Sorkhi, M., Jahed-Motlagh, M. R., Minaei-Bidgoli, B. and Daliri, M. R., "Hybrid Fuzzy Deep Neural Network Toward Temporal-spatial-frequency Features Learning of Motor Imagery Signals," *Sci. Rep.*, **12**, 22334(2022).
13. C, A. *et al.*, "Noise Reduction in CT Images Using a Selective Mean Filter," *Journal of Biomedical Physics & Engineering*, **10**, 623-634(2020).
14. Huang, T., Yang, G. and Tang, G., "A Fast Two-dimensional Median Filtering Algorithm," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, **27**, 13-18(1979).
15. Haddad, R. A. and Akansu, A. N., "A Class of Fast Gaussian Binomial Filters for Speech and Image Processing," *IEEE Transactions on Signal Processing*, **39**, 723-727(1991).
16. Lee, H., "Analysis of Preconcentration Dynamics inside Dead-end Microchannel," *Korean Chem. Eng. Res.*, **61**, 155-161(2023).
17. Kim, S. J., Wang, Y.-C., Lee, J. H., Jang, H. and Han, J., "Concentration Polarization and Nonlinear Electrokinetic Flow near a Nanofluidic Channel," *Phys. Rev. Lett.*, **99**, 044501(2007).
18. Kim, S. J., Li, L. D. and Han, J., "Amplified Electrokinetic Response by Concentration Polarization near Nanofluidic Channel," *Langmuir*, **25**, 7759-7765(2009).
19. Kim, J., Kim, H.-Y., Lee, H. and Kim, S. J., "Pseudo 1-D Micro/Nanofluidic Device for Exact Electrokinetic Responses," *Langmuir*, **32**, 6478-6485(2016).
20. Choi, J. *et al.*, "Selective Preconcentration and Online Collection of Charged Molecules Using Ion Concentration Polarization," *RSC Adv.*, **5**, 66178-66184(2015).
21. Choi, J. *et al.*, "Nanoelectrokinetic Selective Preconcentration Based on Ion Concentration Polarization," *BIOCHIP J.*, **14**, 100-109(2020).
22. Kim, J., Cho, I., Lee, H. and Kim, S. J., "Ion Concentration Polarization by Bifurcated Current Path," *Sci. Rep.*, **7**, 5091(2017).
23. Dydek, E. V. and Bazant, M. Z., "Nonlinear Dynamics of Ion Concentration Polarization in Porous Media: The Leaky Membrane Model," *AIChE Journal*, **59**, 3539-3555(2013).
24. Robertson, R. M., Laib, S. and Smith, D. E., "Diffusion of Isolated DNA Molecules: Dependence on Length and Topology," *Proc. Natl. Acad. Sci. U.S.A.*, **103**, 7310-7314(2006).
25. Salieb-Beugelaar, G. B., Dorfman, K. D., van den Berg, A. and Eijkel, J. C. T., "Electrophoretic Separation of DNA in Gels and Nanostructures," *Lab Chip*, **9**, 2508-2523(2009).
26. Yap, K. K., Fukuda, K., Vail, J. R., Wong, J. and Masen, M. A., "Spatiotemporal Mapping for In-situ and Real-time Tribological Analysis in Polymer-metal Contacts," *Tribology International*, **171**, 107533(2022).
27. Posner, J. D., Pérez, C. L. and Santiago, J. G., "Electric Fields Yield Chaos in Microflows," *Proc. Natl. Acad. Sci. U.S.A.*, **109**, 14353-14356(2012).
28. Kwak, R., Pham, V. S. and Han, J., "Sheltering the Perturbed Vortical Layer of Electroconvection Under Shear Flow," *J. Fluid Mech.*, **813**, 799-823(2017).
29. Cho, S.-Y. *et al.*, "Finding Hidden Signals in Chemical Sensors Using Deep Learning," *Anal. Chem.*, **92**, 6529-6537(2020).

Authors

Gyeonghee Kang: Undergraduate Student, Department of Chemical Engineering, Jeju National Univeristy, Jeju 63243, Korea; rudgml4336@naver.com
Minji Kim: Undergraduate Student, Department of Chemical Engineering, Jeju National Univeristy, Jeju 63243, Korea; kjkk6672@naver.com
Hyomin Lee: Assicoate Professor, Department of Chemical Engineering, Jeju National University, Jeju 63243, Korea; fluid@jejunu.ac.kr