УДК 539.261

© 2008 р. С.В. Баловсяк, І.М. Фодчук

Чернівецький національний університет ім. Ю.Федьковича, Чернівці

ВИКОРИСТАННЯ ШТУЧНИХ НЕЙРОННИХ МЕРЕЖ ДЛЯ ВИЗНАЧЕННЯ ПАРАМЕТРІВ НАПІВПРОВІДНИКІВ ЗА ДАНИМИ *X*-ПРОМЕНЕВИХ МЕТОДІВ

Подано результати визначення структурних параметрів кристалів CdTe за допомогою штучних нейронних мереж на основі *X*-променевих кривих дифракційного відбивання. Використано тришарову мережу зі зворотним розповсюдженням помилки та з динамічним додаванням нейронів. За допомогою створеного програмного забезпечення проведено моделювання роботи нейромережі. Проаналізовано вплив параметрів нейронної мережі на час навчання і на помилку мережі для тестових зразків. Показано, що динамічне додавання нейронів у поєднанні з коректним вибором топології й параметрів мережі дозволяє зменшити час навчання.

The results of determination of structural parameters of the CdTe crystals by artificial neural networks on the basis of x-ray diffraction curves are presented. The three-layered network with back-propagation of errors and with dynamic addition of neurons is used. By the created software the neural networks simulation are realized. Influencing of parameters of neural network in a time of learning and on the errors of network for test samples is analyzed. It is shown that dynamic additions of neurons in combination with the correct choice topologies and parameters of network allows to decrease learning time.

Вступ

Для розв'язання багатьох практичних задач потрібно визначати параметри напівпровідникових матеріалів, зокрема характеристики їх структурної досконалості, на основі даних *X*-променевих методів. До найперспективніших *X*-променевих методів належать високороздільна дифрактометрія та рефлектометрія в умовах повного зовнішнього відбивання Х-променів [1]. У загальному випадку залежність розподілу інтенсивності *X*-променевих кривих від структурних параметрів зразків дуже складна й неоднозначна, що значно ускладнює розв'язання оберненої задачі (відновлення структурних параметрів). Водночас для розв'язання подібних складних задач, наприклад у сфері медичної й технічної діагностики, ефективно використовується новий метод – штучні нейронні мережі [2-4]. Тому метою даної роботи є вивчення можливостей використання штучних нейронних мереж для знаходження структурних параметрів зразків на основі даних *X*-променевих методів. При цьому вид і топологію штучних нейронних мереж вибрано відповідно до особливостей вхідних і вихідних даних *X*-променевих методів.

1. Теоретична частина

Штучні нейронні мережі (*neural networks*) – надзвичайно спрощена модель біологічних нейронних мереж. Особливістю нейромереж (НМ) є те, що вони навчаються, а не програмуються. Існують програмні й апаратні моделі НМ, але на даний час звичайно виконують програмну реалізацію нейромереж. Виділяють наступні режими роботи НМ:

1. Навчання (відомі вхідні й вихідні дані, визначаються вагові коефіцієнти).

2. Тестування (відомі вхідні й вихідні дані, порівнюються розраховані вихідні дані з істинними).

3. Діагностика (реальне визначення результатів за вхідними даними).

Основними компонентами НМ є нейрони (*neurons*), які з'єднані зваженими зв'язками.

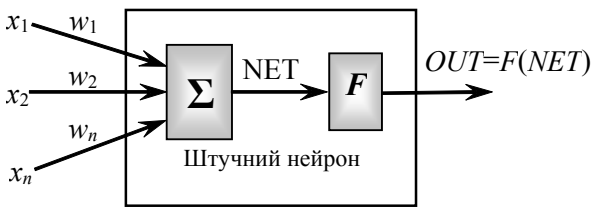

Рис.1. Штучний нейрон з активаційною функцією *F*

З кожним зв'язком між нейронами пов'язаний ваговий коефіцієнт (*weighting coefficient*). На вхід штучного нейрона поступає множина сигналів x_1, x_2,…, x_n (вектор X), які є виходами інших нейронів. Кожен сигнал множиться на відповідну вагу w_1, w_2,…, w_n (вектор W) і всі виходи підсумовуються у блоці Σ, визначаючи рівень активації нейрона (рис. 1). Блок Σ складає зважені входи, алгебраїчно створюючи вихід NET:

$$NET = X\,W. \qquad (1)$$

1.1. Активаційні функції

Сигнал NET далі, як правило, перетворюється активаційною функцією F і дає вихідний нейронний сигнал $OUT = F(NET)$. Активаційна функція $F(NET)$ може бути пороговою бінарною, лінійною обмеженою, функцією гіперболічного тангенса, але найбільш поширена сигмоїдна (S-подібна або логістична) функція (рис. 2)

$$OUT = \frac{1}{1+e^{-C\cdot NET}}, \qquad (2)$$

де C – константа (зазвичай C=1)

З виразу (2) для сигмоїда очевидно, що вихідне значення нейрона лежить у діапазоні [0,1]. Популярність сигмоїдної функції зумовлена такими її властивості:

1) здатність підсилювати слабкі сигнали сильніше, ніж великі, і опиратися "насиченню" від потужних сигналів;

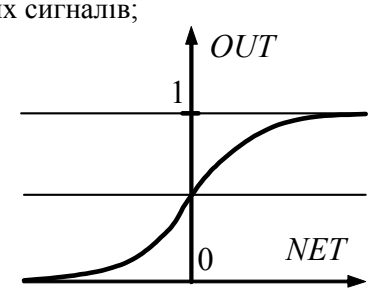

Рис. 2. Сигмоїдна функція активації

2) монотонність і диференційованість на всій осі абсцис;

3) простий вираз для похідної

$$F'(NET) = C\cdot F(NET)\cdot(1-F(NET)), \qquad (3)$$

що дає можливість використовувати широкий спектр оптимізаційних алгоритмів.

1.2. Багатошарові нейронні мережі

Хоча один нейрон здатний виконувати прості процедури розпізнавання, сила нейронних обчислень виникає від з'єднань нейронів у мережах. Багатошарові мережі (рис. 3) мають значно більші можливості, ніж одношарові. Проте багатошарові мережі можуть привести до збільшення обчислювальної потужності порівняно з одношаровими лише в тому випадку, якщо активаційна функція між шарами буде нелінійною.

У роботі використано тришарову нейронну мережу, оскільки при меншій кількості шарів можливості мережі суттєво обмежені, а при більшій кількості шарів значно збільшується час розрахунку. Для навчання мережі використано алгоритм зворотного розповсюдження помилки.

Розглянемо структуру 3-шарової нейромережі (рис. 3).

1. Вхідний шар описується вектором $X=\{x_1,..., x_i,...,x_{QX}\}$. Розмір навчальної множини (кількість векторів X) дорівнює QN, номер вектора $n=\overline{1,Q_N}$.

2. Приховані шари:

Перший шар (рівень L=1) описується вектором значень нейронів $V^1=\{v_1^1,..., v_{k1}^1,..., v_{QV1}^1\}$ та вектором різниці $D^1=\{d_1^1,..., d_{k1}^1,..., d_{QV1}^1\}$, зв'язки вхідного шару з наступним визначаються матрицею вагових коефіцієнтів W^1.

Другий шар (рівень L=2) описується вектором

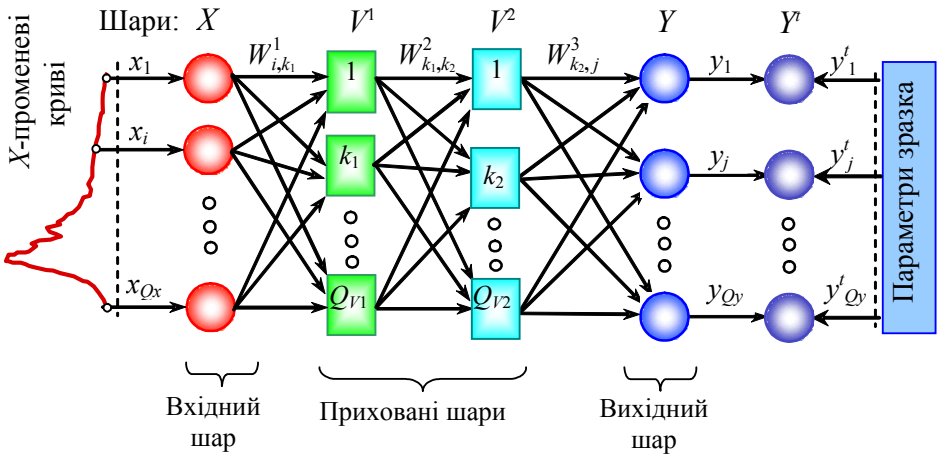

Рис. 3. Тришарова НМ

значень нейронів $V^2=\{\nu_1^2,...,\nu_{k2}^2,...,\nu_{OV2}^2\}$ та вектором різниці $D^2\{d_1^2,...,d_{k2}^2,...,d_{OV2}^2\}\}$, зв'язки першого шару з наступним визначаються матрицею вагових коефіцієнтів W^2.

3. Вихідний шар (рівень L=3) описується вектором дійсних значень нейронів $Y=\{y_1,..,y_j,...,y_{QY}\}$, вектором істинних значень нейронів $Y^T=\{y_1^T,...,y_j^T,...,y_{OY}^T\}$ та вектором різниці $D^3=\{d_1^3,..,d_j^3,...,d_{OY}^3\}$, зв'язки другого шару з наступним визначаються матрицею вагових коефіцієнтів W^3. Розмір навчальної вибірки (кількість векторів Y^T) дорівнює Q_M, номер вектора $m=\overline{1,Q_M}$.

1.3. Алгоритм зворотного розповсюдження помилки

Зворотне розповсюдження помилки (*back-propagation*) означає, що сигнали помилки з виходу мережі використовуються для корекції ваги попередніх шарів. Навчання нейромережі відбувається за наступним алгоритмом (рис. 4), де кожна ітерація процесу навчання називається епохою. Процес навчання завершується, якщо кількість епох e перевищує допустиму Q_E або похибка мережі ε_k менша допустимої ε_{kMin}.

Виділяють такі етапи навчання НМ.

1. Ініціалізація. Початкові значення матриці вагових коефіцієнтів W приймаються такими, що дорівнюють малим випадковим значенням, наприклад:

$$W_{i,k1}^1=(Rnd-0{,}5)\cdot 2\Delta W, \qquad (4)$$

де $i=\overline{1,Q_X}$, $k_1=\overline{1,Q_{V1}}$, Rnd – значення рівномірно розподіленої випадкової величини в діапазоні [0, 1], ΔW=0,3.

У вагових матрицях рядки відповідають елементам, від яких йдуть зв'язки, а стовпці – до яких йдуть зв'язки.

2. Нормалізація (масштабування) початкових значень усіх векторів X, Y^T (для кожного типу даних вектора окремо) в діапазон ($MinN$, $MaxN$), наприклад:

$$X_i=MinN+\frac{(X_i^p-X_{\min})\cdot(MaxN-MinN)}{(X_{\max}-X_{\min})}, \qquad (5)$$

де $MinN$=0,1, $MaxN$=0,9, x_i^p – значення елемента вектора до нормалізації.

3. Пряме розповсюдження полягає у знаходженні вихідного вектора Y на основі вхідного X за наступними формулами.

Шар 1:

$$Net_{k1}=\sum_{i=1}^{Q_X}X_i\cdot W_{i,k1}^1,$$

$$V_{k1}^1=f\left(\frac{1}{1+\exp(-Net_{k1})}\right), \qquad (6)$$

де $k_1=\overline{1,Q_{V1}}$.

Шар 2:

$$Net_{k2}=\sum_{k1=1}^{Q_{V1}}V^1{}_{k1}\cdot W_{k1,k2}^2,$$

$$V_{k2}^2=f\left(\frac{1}{1+\exp(-Net_{k2})}\right), \qquad (7)$$

де $k_2=\overline{1,Q_{V2}}$.

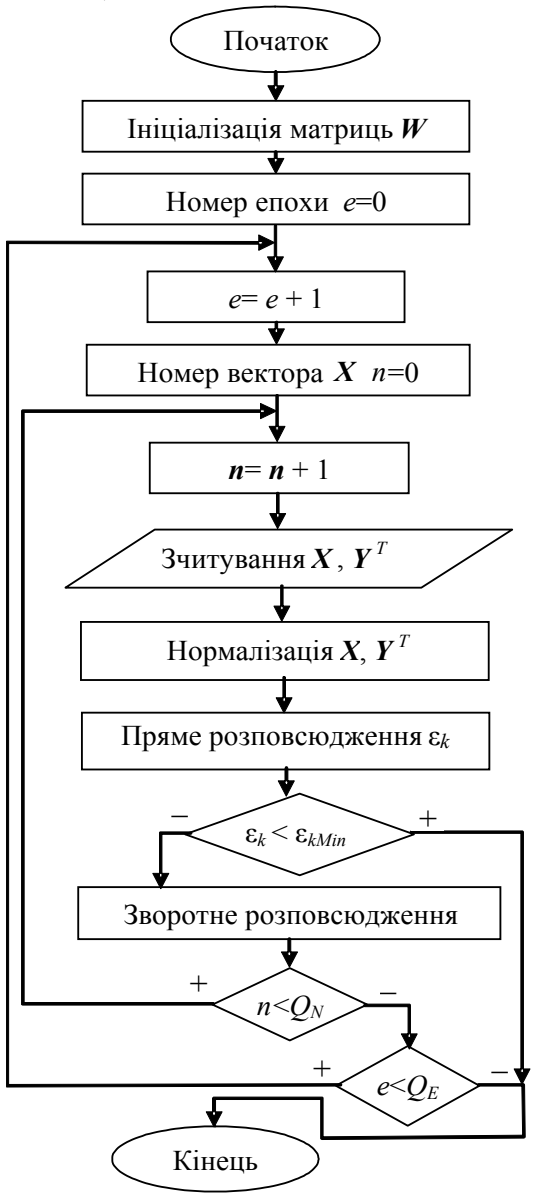

Рис. 4. Алгоритм навчання нейромережі зі зворотним розповсюдженням помилки

Шар 3:

$$Net_j = \sum_{j=1}^{Q_Y} V_{k2}^2 \cdot W_{k2,j}^3,$$

$$Y_j = f\left(\frac{1}{1+\exp(-Net_j)}\right), \qquad (8)$$

де $j = \overline{1, Q_Y}$

У результаті прямого розповсюдження можна обчислити похибки навчання мережі:

$\varepsilon_{Ln} = \sum_{j=1}^{Q_Y} \left| Y_j - Y_j^T \right|$ – лінійна похибка для вектора з номером n,

$\varepsilon_L = \sum_{n=1}^{Q_N} \sum_{j=1}^{Q_Y} \left| Y_{j,n} - Y_{j,n}^T \right|$ – лінійна похибка для всіх векторів навчальної множини,

$\varepsilon_{kn} = \frac{1}{2} \sum_{j=1}^{Q_Y} \left(Y_{j,n} - Y_{j,n}^T \right)^2$ – квадратична похибка для одного вектора з номером n,

$\varepsilon_k = \frac{1}{2} \sum_{n=1}^{Q_N} \sum_{j=1}^{Q_Y} \left(Y_{j,n} - Y_{j,n}^T \right)^2$ – сумарна квадратична похибка для всіх векторів навчальної множини,

$$\varepsilon_\sigma = \frac{1}{Q_N Q_Y} \sum_{n=1}^{Q_N} \sum_{j=1}^{Q_Y} \left(\frac{\left| Y_{j,n} - Y_{j,n}^T \right|}{\left| Y_{j,n}^T \right| + 1} \right) \cdot 100\%$$ – середня відносна похибка.

4. Зворотне розповсюдження похибки полягає в корекції вагових коефіцієнтів через сигнал різниці D.

Шар 3:

$$D_j^3 = Y_j \cdot (1 - Y_j)(Y_j^T - Y_j),$$

$$W_{k2,j}^{3(e)} = W_{k2,j}^{3(e-1)} + \eta_Y \cdot D_j^3 \cdot V_{k2}^2, \qquad (9)$$

де $j = \overline{1, Q_Y}$, e – номер епохи. Оскільки як активаційна функція використовується сигмоїдна, то різниця векторів $Y^T - Y$ множиться на похідну від сигмоїдної функції: $Y(1-Y)$.

Шар 2:

$$D_{k_2}^2 = \sum_{j}^{Q_Y} V_{k_2}^2 \cdot (1 - V_{k_2}^2) \cdot (D_j^3 \cdot W_{k_2,j}^3),$$

$$W_{k1,k2}^{2(e)} = W_{k1,k2}^{2(e-1)} + \eta_{L2} \cdot D_{k2}^2 \cdot V_{k1}^1, \qquad (10)$$

де $k_2 = \overline{1, Q_{V2}}$

Шар 1:

$$D_{k1}^1 = \sum_{k2}^{Q_{V2}} V_{k1}^1 \cdot (1 - V_{k1}^1) \cdot (D_{k2}^2 \cdot W_{k1,k2}^2),$$

$$W_{i,k1}^{1(e)} = W_{i,k1}^{1(e-1)} + \eta_{L1} \cdot D_{k1}^1 \cdot X_i, \qquad (11)$$

де $k_1 = \overline{1, Q_{V1}}$, η_Y, η_{L2}, η_{L1} – норми навчання (значення норм навчання, наприклад, 0,2).

2. Початкові дані

Як об'єкти дослідження (зразки CdTe №1, №2, №3) використано епітаксійні шари $Cd_xHg_{1-x}Te$ (111) (x=0,252), нарощені на нелегованих підкладках CdTe (111) [5]. Зразки №2 та №3 імплантовані іонами As з енергією E=100 кеВ. Для зразка №2 доза імплантованих іонів D=10^{15}см$^{-2}$, а для №3 доза D=1,2·10^{15}см$^{-2}$. Для двох областей кожного зразка визначено їх структурні характеристики: товщину області аморфізації, значення густини дислокацій у припущенні їх хаотичного розподілу, глибину максимальної деформації, значення максимальної деформації (таблиця 1).

Усі зразки досліджені методом X-променевої дифрактометрії, у результаті чого для кожного зразка отримано кілька експериментальних кривих дифракційного відбивання (рис. 5). Як вхідні дані для НМ (вектор X) використано значення інтенсивностей X-променевих кривих в Q_X точках, а як вихідні дані (вектор Y^T) – структурні параметри зразків, а також енергію й дозу імплантованих іонів. Для навчання мережі використано вибірку, яка містила криві дифракційного відбивання й структурні параметри для двох областей зразків CdTe №1-3 (таблиця 1).

Таблиця 1. Структурні характеристики зразків CdTe [5]

Зразок	Область	Товщина області аморфізації z_A, мкм	Густина дислокацій n_n, 10^7 см$^{-2}$	Глибина максимальної деформації z_{max}, мкм	Максимальна деформація $\Delta d_{max}/d$, 10^{-3}
№1	1	0	0,068	0	0
	2	0	0,05	0	0
№2	1	0,3	0,094	0,14	0,06
	2	0,3	0,094	0,14	0,06
№3	1	0,38	1,2	0,16	0,5
	2	0,38	0,6	0,16	0,4

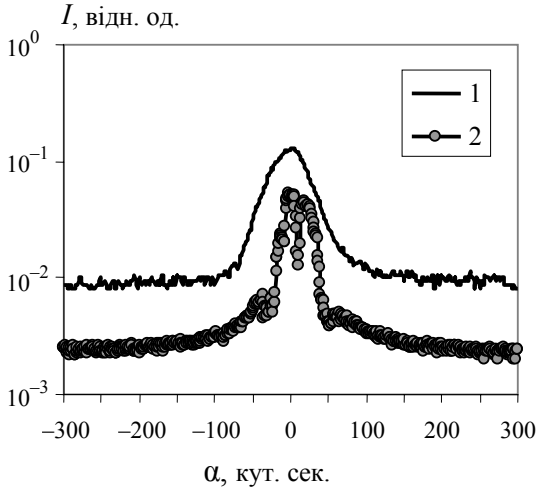

Рис. 5. Криві дифракційного відбивання для зразків CdTe, відбивання (333), CuK_α-випромінювання: зразок №1 (область 1), №2 (область 1)

3. Програмне забезпечення

Програма для моделювання НМ створена в *Delphi* 6, яка також забезпечує нормалізацію вхідних і вихідних даних (рис. 6). Програма виконує навчання 3-шарової нейромережі методом зворотного розповсюдження помилки. На вхідний шар *X* подаються розподіли інтенсивностей *X*-променевих кривих у вигляді одномірних векторів. На основі значень вектора *X* розраховуються вектори прихованих шарів V^1, V^2 та вихідного шару *Y*. При навчанні мережі отримані виходи *Y* порівнюються з істинними Y^T, у випадку їх відмінності відбувається корекція матриць вагових коефіцієнтів W^1, W^2, W^3. Процес навчання завершується, якщо отримана сумарна квадратична похибка ε_k менша, ніж задана. Навчальна й тестова вибірки зберігаються у базі даних програми.

При навчанні мережі на основі експериментальних кривих (кількість точок Q_X=128, похибка ε_{kMin}=0,001, кількість нейронів у прихованих шарах Q_{V1}=160, Q_{V1}=80) час навчання виявився значним (таблиця 2). Для розв'язання цієї проблеми використано динамічне додавання нейронів у вхідний шар *X*. На початку навчання (рівень розмірності першої вагової матриці *s*=1) в шарі *X* було 8 нейронів (які описували 8 фрагментів кривої), а в процесі навчання на кожному рівні *s* кількість нейронів у шарі *X* подвоювалася (на останньому етапі один нейрон відповідав одній точці кривої).

Обчислення вагової матриці наступного рівня *s* (більшої розмірності) виконувалося за формулою

$$W_{2i,k1}^{1(s)} = W_{i,k1}^{1(s-1)} \cdot k_W , \qquad (12)$$

де коефіцієнт k_W=1,5, $i = \overline{1, Q_X^{(s-1)}}$, $k_1 = \overline{1, Q_{V1}}$, $Q_X^{(s-1)}$ – кількість нейронів шару *X* для рівня (*s*-1).

$$W_{i,k1}^{1(s)} = W_{i,k1}^{1(s)} + Rnd \cdot k_{Rnd} , \qquad (13)$$

де коефіцієнт k_{Rnd}=2, $i = \overline{1, Q_X^{(s)}}$, *Rnd* – значення рівномірно розподіленої випадкової величини в діапазоні [0, 1].

Значення коефіцієнтів k_W, k_{Rnd} вибрано за умови мінімізації часу навчання НМ. У результаті послідовне збільшення кількості нейронів значно прискорило навчання НМ (таблиця 2). Отже, на початкових рівнях *s* мережа навчається на *X*-променевих кривих спрощеної форми (менша кількість точок), а на останньому рівні *s* використовуються *X*-променеві криві максимальної розмірності. Такий принцип обробки інформації, коли спочатку використовується спрощена й узагальнена модель об'єкта, яка поступово ускладнюється й деталізується, досить ефективний при обробці великих обсягів даних. Варто зауважити, що подібний принцип поступової деталізації використовується в зоровій системі людини та деяких системах технічного зору [6].

Тестування нейронної мережі проведено за допомогою контрольної вибірки кривих гойдання (таблиця 3, рис. 7). У таблиці 3 вектором Y^T позначено відомі параметри зразків, а *Y* – визначені за допомогою НМ. Отримана похибки для тестових векторів (CdTe №2 (область 3) – ε_k= 0,015, CdTe №3 (область3) – ε_k=0,054) перевищує помилку навчання мережі тільки на порядок, що свідчить про задовільне відновлення структурних характеристик НМ за допомогою НМ (таблиця 3).

Висновки

У результаті проведених досліджень створено програмну модель нейтронної мережі, призначену для визначення структурних характеристик кристалів на основі їх кривих дифракційного відбивання. Проведено оптимізацію топології та параметрів навчання НМ з урахуванням специфіки *X*-променевих кривих, наприклад кількості шарів і норм навчання. Показано, що зменшення часу навчання НМ можна досягти завдяки динамічному додаванню нейронів, тобто спочатку виконується обробка спрощеного розподілу інтенсивності *X*-променевих кривих, а потім – більш детального.

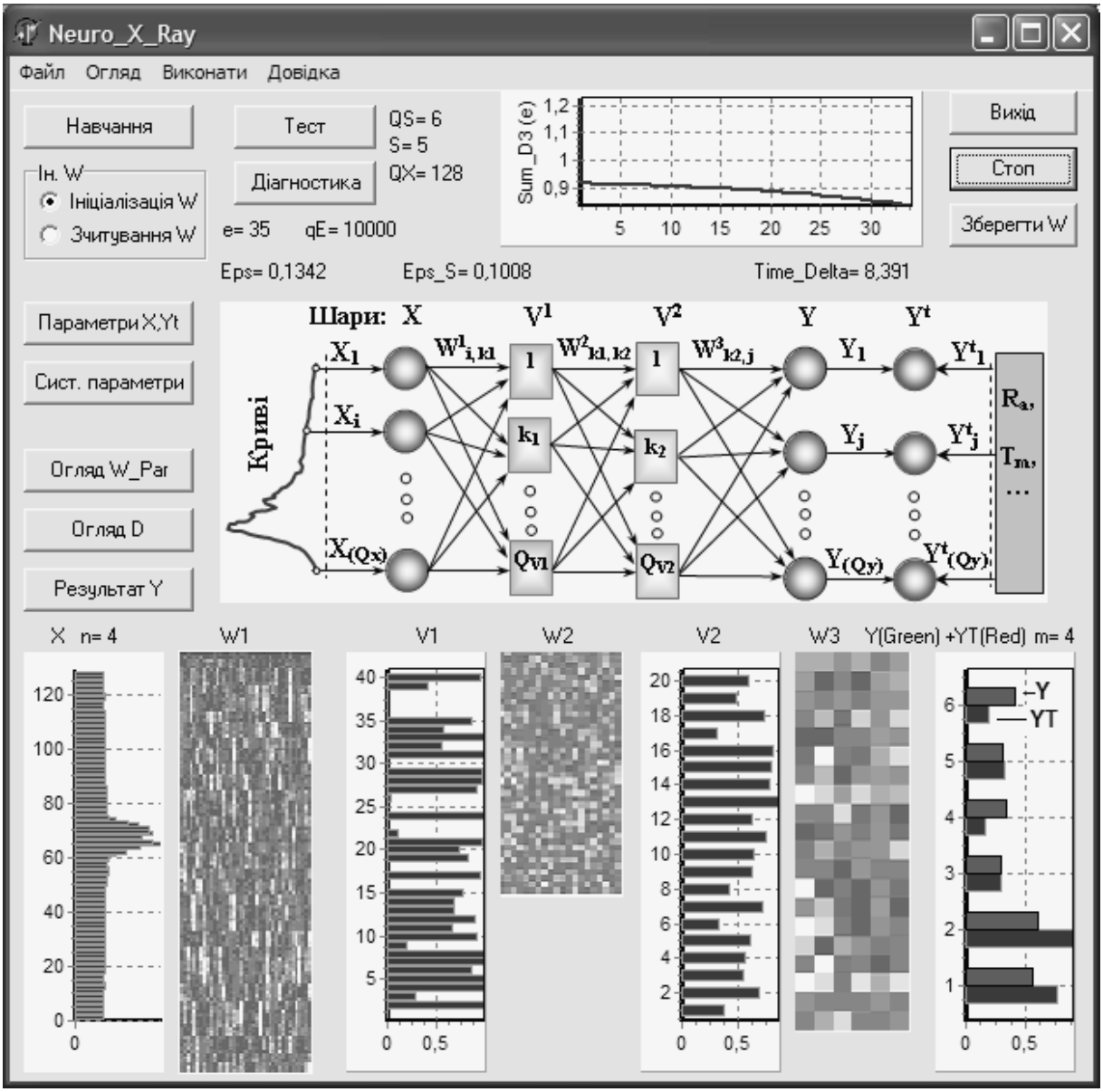

Рис. 6. Інтерфейс програми для моделювання НМ

Таблиця 2. Параметри навчання НМ для статичного (без зміни кількості нейронів) і динамічного режимів

Кількість нейронів		Режим навчання		Режим навчання	
Q_{V1}	Q_{V2}	статичний	динамічний	статичний	динамічний
		Час навчання, с		Кількість епох	
40	20	10,089	4,612	10835	4769
80	40	16,917	3,951	8063	1743
160	80	33,701	8,243	4329	981

Таблиця 3. Структурні характеристики зразків CdTe, отримані за допомогою НМ

Зразок	Область	Вектор	Товщина області аморфізації z_A, мкм	Густина дислокацій n_n, 10^7 см$^{-2}$	Глибина максимальної деформації z_{max}, мкм	Максимальна деформація $\Delta d_{max}/d$, 10^{-3}
№2	3	Y^T	0,3	0,094	0,14	0,06
		Y	0,293	0,048	0,136	0,042
№3	3	Y^T	0,38	1,2	0,16	0,5
		Y	0,337	0,872	0,139	0,467

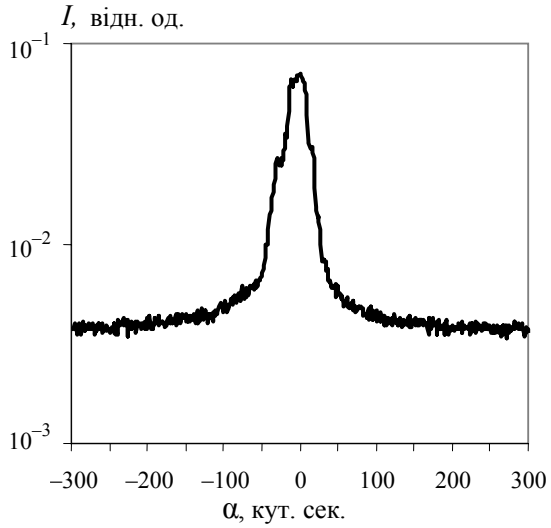

Рис. 7. Крива гойдання для зразка CdTe №2 (область 3)

Новизна роботи полягає у використанні НМ для обробки кривих дифракційного відбивання для зразків CdTe, а також у реалізації динамічного додавання нейронів до шарів НМ.

Отже, штучні нейронні мережі відкривають нові можливості для розв'язання складних і неоднозначних задач, зокрема обернених задач відновлення структурних параметрів зразків на основі експериментальних *X*-променевих кривих. Значною перевагою НМ з точки зору трудомісткості є те, що вони навчаються, а не програмуються. Використання НМ особливо ефективне в поєднанні з класичними методами аналізу *X*-променевих кривих, що дозволяє оцінити коректність роботи НМ.

СПИСОК ЛІТЕРАТУРИ

1. *Афанасьев А.М., Александров П.А., Имамов Р.М.* Рентгенодифракционная диагностика субмикронных слоев. – М.: Наука, 1980.
2. *Заенцев И.В*. Нейронные сети: основные модели. – Воронеж: Изд. Воронеж, ун-та, 1999.
3. *Рутковская Д., Пилиньский М., Рутковский Л.* Нейронные сети, генетические алгоритмы и нечеткие системы. – М.: Горячая линия - Телеком, 2004.
4. *Уосермен Ф.* Нейрокомпьютерная техника. Теория и практика. – 1992.
5. *Заплитный Р.А., Каземирский Т.А., Фодчук И.М., Свянтек З.* Структурные изменения в эпитаксиальных структурах, модифицированных ионной имплантацией // Металлофизика и новейшие технологии. – 2006. – **27**, №8. – С. 915-932.
6. *Кутковецький В.Я*. Розпізнавання образів: Навчальний посібник. – Миколаїв: Вид-во МДГУ ім. П. Могили, 2003.