Технічні науки

УДК 004.89

Єфремова Софія Валеріївна

студентка

Національного технічного університету України

«Київський політехнічний інститут імені Ігоря Сікорського»

Ефремова София Валерьевна

студентка

Национального технического университета Украины

«Киевский политехнический институт имени Игоря Сикорского»

Yefremova Sofiia

Student of the

National Technical University of Ukraine

«Igor Sikorsky Kyiv Polytechnic Institute»

Науковий керівник:

Носовець Олена Костянтинівна

кандидат технічних наук, доцент кафедри біомедичної кібернетики

Національний технічний університет України

«Київський політехнічний інститут імені Ігоря Сікорського»

ПРОГРАМНИЙ ДОДАТОК ДЛЯ ПРОГНОЗУВАННЯ ХВОРОБИ АЛЬЦГЕЙМЕР НА ОСНОВІ РІВНЯ ЕКСПРЕСІЇ ГЕНІВ

ПРОГРАММНОЕ ПРИЛОЖЕНИЕ ДЛЯ ПРОГНОЗИРОВАНИЯ БОЛЕЗНИ АЛЬЦГЕЙМЕР НА ОСНОВЕ УРОВНЯ ЭКСПРЕССИИ ГЕНОВ

A SOFTWARE APPLICATION FOR PREDICTING ALZHEIMER'S DISEASE BASED ON THE LEVEL OF GENE EXPRESSION

Анотація. *Дана стаття розглядає процес створення програмного додатку, який виконує прогнозування хвороби Альцгеймер на основі даних експресії генів здорових та хворих пацієнтів. Об'єкт розгляду дослідження є зразки експресії генів, взятих у результаті дослідження, в якому використовували сторону середньої скроневої звивини мозку заморожених зразків.*

Ключові слова: *рівень експресії генів, нейрона мережа, машинне навчання, хвороба Альцгеймера, глибоке навчання, експресія генів, методи відбору головних ознак.*

Аннотация. *Данная статья рассматривает процесс создания программного приложения, которое выполняет прогнозирования болезни Альцгеймер на основе данных экспрессии генов здоровых и больных пациентов. Объект рассмотрения исследования является образцы экспрессии генов, взятых в результате исследования, в котором использовали сторону средней височной извилины мозга замороженных образцов.*

Ключевые слова: *уровень экспрессии генов, нейронная сеть, машинное обучение, болезнь Альцгеймера, глубокое обучение, экспрессия генов, методы отбора главных признаков.*

Summary. *This article discusses the process of creating a software application that predicts Alzheimer's disease based on gene expression data in healthy and sick patients. The object of the study is the expression samples of genes taken from the study, which used the side of the middle temporal gyrus of the brain of frozen samples*

Key words: *Gene expression level, neural network, machine learning, Alzheimer 's disease, deep learning, gene expression, methods of selection of main features.*

Постановка проблеми. На даний момент у світі біля 50-ти мільйонів чоловік страждають на хворобу Альцгеймера. Надалі очікується, що ця недуга, що є найбільш поширеною формою деменції, буде рости в міру старіння населення. Крім безпосереднього впливу на здоров'я і благополуччя людини, довгостроковий догляд порушених осіб накладає значне економічне тягар. Хвороба Альцгеймера швидко перетворюється в критичну глобальну проблему для здоров'я і економіки. Саме тому існує потреба в зниженні ризиків за рахунок прийняття необхідних запобіжних заходів для поширення її впливу шляхом діагностики людей на ранніх стадіях. Прогнозування стану захворювання людини на складну хворобу за допомогою вимірювання експресії генів є важливою частиною багатьох наук про життя, оскільки можливість кількісно оцінити рівень експресії конкретного гена в клітині, тканині або організмі може надати багато цінної інформації. Вимір експресії генів може визначити схильність людини до хвороби.

Метою даної роботи є прогнозування хвороби на основі даних експресії генів здорових та хворих на Альцгеймер пацієнтів.

Виклад основного матеріалу. База даних для дипломної роботи взято з загальнодоступного ресурсу Національного Центру Біотехнологічної Інформації (NCBI GEO) [3]. Аналіз проводиться з використанням зразків експресії гена NCBI, взятих у результаті дослідження, в якому використовували сторону середньої скроневої звивини мозку заморожених зразків. Загальною проблемою при роботі з генетичними даними велика кількість ознак (набору генів) для обмеженої кількості спостережень (пацієнтів). Щоб уникнути цієї проблеми було використано методи вибору ключових (зразки з найбільшим рівнем експресії) ознак. Початкова база даних складається з 12850 тис. наборів ознак та 180 спостережень. Спостереження поділяються на два класи : здорові пацієнти (AD Healthy) та пацієнти з хворобою Альцгеймер (AD).

Вибір ознак є важливою операцією при обробці генетичних даних. Гени з найвищім рівнем експресії (ключові) збільшують наше розуміння механізму формування хвороби і дозволяють прогнозувати потенційну небезпека ураження. Застосування методів вибору ознак дозволяє визначити невелику кількість важливих генів, які можуть бути використані як біомаркери відповідної хвороби. У даній роботі розглянуто 3 методи вибору ознак такі як: Дискримінантний аналіз Фішера, Двовибірковий t-критерій Ст'юдента та Кореляційний аналіз.

У підході Фішера найбільша вага призначається ознаці яка характеризується великою різницею середнього значення у двох вивчених класах і невелике значення стандартного відхилення в межах кожного класу. Дискримінаційну здатність ознаки f визначається у вигляді [3]:

$$S_{12}(f) = \frac{|c_1 - c_2|}{\sigma_1 + \sigma_2}, \quad (1)$$

де: c1 і c2 – представляють середні значення для класів 1 та 2 відповідно, тоді як σ1 та σ2 є відповідними стандартним відхилення.

Велике значення S12 (f) вказує на хороший дискримінаційну здатність ознаки класу. З іншого боку маленький значення є показником незначності ознаки у визнанні цих двох класів.

Наступним використовуваним методом відбору є двовибірковий t-критерій. Двовибірковий t-критерій є одним із найбільш часто використовуваних тестів гіпотез. Застосовується для порівняння, чи справді середня різниця між двома групами є дійсно значною, чи це пов'язано із випадковим випадком [5]. Одним з головних факторів використання t-критерію полягає в тому що, всі дані з вибірки, що досліджується, повинні слідувати нормальному розподілу. Сутність нульової гіпотези теза t- критерію полягає в тому, що дані в рівності математичних очікувань класів 1 і 2. t- статистика для перевірки гіпотези дорівнює:

$$t = \frac{c_1 - c_2}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}}, \quad (2)$$

де n і m – розміри вибірки обох класів [2] .

Кореляційний аналіз

Для Дискримінантне значення функції S (f) для розпізнавання одного класу з інших K класів визначається наступним чином [4]:

$$S(f) = \frac{\sum_k^K P_k (m_{k-m})^2}{\sigma^2(f) \sum_{k=1}^K P_k (1 - P_k)}, \quad (3)$$

де m – середнє значення ознаки для всіх даних,

mk – середнє значення ознаки для даних k-го класу, σ2

(f) – дисперсія ознаки

Pk – це ймовірність k-го класу виникнення в наборі даних (передбачається рівномірний розподіл).

У цій роботі кількість класів становить 2 (K = 2). При рівномірному розподілі обох класів попереднє рівняння можна спростити до наступної формули:

$$S_{12}(f) = \frac{(m_1(f) - m(f))^2 + (m_2(f) - m(f))^2}{2\sigma^2(f)}, \quad (4)$$

Велике значення S12 (f) вказує на хорошу дискримінаційну здатність ознака для розпізнавання двох класів.

Методи відбору ознак, описані в попередніх розділах, застосовувались для отримання порядку генів, відсортованих за зменшенням. В результаті цих експериментів було відібрано три підмножини із 100 найбільш генів з найбільшим рівнем екпресії.

Були застосовані такі скорочення:

- ДА – дискримінантний аналіз Фішера;
- КА – Кореляційний аналіз;
- t-критерій – Двовибірковий t-критерій Ст'юдента.

Як і слід було очікувати, методи відібрали різні набори генів. У таблиці 1 показано, скільки однакових генів серед перших 100

найважливіших було відібрано різними методами.

Таблиця 1

Відсоток збігу між методами відбору даних

	ДА	t-критерій	КА
ДА	100	36	90
t-критерій	36	100	66
КА	90	66	100

Зміст вибраних наборів відрізняється від методу до методу. Аналізуючи їх, ми можемо виявити, що небагато методів ідентифікували велику кількість однакових генів. Результати, що збігаються між результатами Фішера та кореляцією з методами класу, охоплюють 90% генів. З іншого боку, деякі з них призвели до різних наборів, тобто дискримінантний аналіз та результати t-критерію накладаються лише на 36%. На рис. 1 зображено рівні експресії у всіх пацієнтів для найважливішого гена, обраного методом Дискримінантного аналізу. Як можна побачити, середнє значення спостережень, що належать до класу хвороби Альцгеймера (AD), суттєво відрізняється від класу здорових показників.

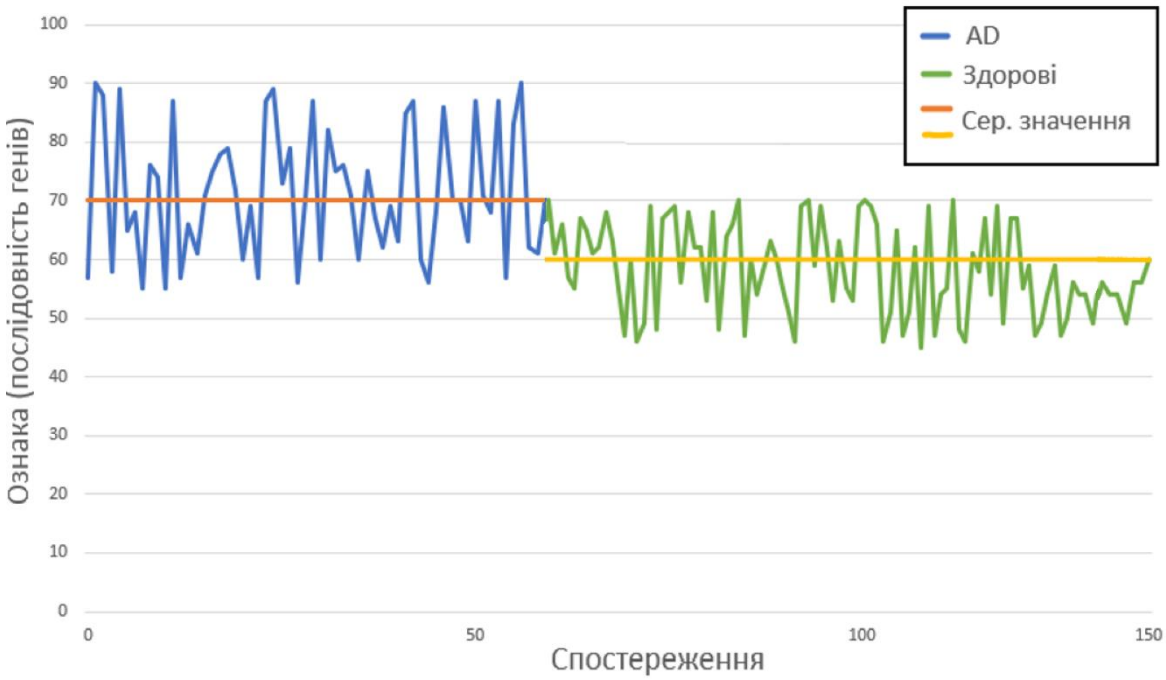

Рис. 1. Рівні експресії гена, що належить до найбільш значущих груп

У результаті відбору ознак було сформовано нову базу, яка складається з 250 наборів ознак та 180 спостережень. структуру якої наведено на рис. 2 Далі було сформовано нову базу, яка складалась з повторюваних ознак, тобто даних, які повторювались хоча б у двох файлах. Кінцевим результатом вважається сформована база, яка складається з 250 наборів ознак та 180 спостережень.

SampleID	ACACB	ACSS1	ADAMTS1	ADARB2	Label_No	Label_AD
SID001	10.277	10.178	8.337	8.523	0	1
SID002	9.305	8.877	8.645	7.368	0	1
SID003	8.73	8.48	8.698	7.554	0	1
SID004	10.521	10.272	8.858	7.777	0	1
SID005	9.383	9.838	10.467	7.687	0	1
SID006	9.678	10.188	8.475	7.653	0	1
SID007	10.692	10.242	7.97	8.093	0	1
SID008	10.022	9.584	8.164	8.776	0	1
SID009	9.659	9.579	8.685	8.498	0	1
SID010	9.293	9.89	8.934	8.322	0	1
SID011	8.52	8.64	7.901	7.919	0	1
SID012	9.452	8.848	7.744	7.358	0	1
SID013	9.878	9.322	8.577	7.36	0	1

Рис. 2. Фрагмент підготовленої БД

У даній роботі використовується модель глибинної нейронної мережі (DNN). Для визначення оптимальних гіперпараметрів запропонованої моделі була використана Байєсівська оптимізація, оскільки вона ефективно розкриває загальні максимуми невідомої функції (чорного ящика), такі як значення точності для набору перевірок у визначеному просторі параметрів. В результаті оптимізації було визначено комбінацію параметрів з найкращою точністю до параметрів які можна побачити на слайді:

- кількість прихованих шарів 8;
- кількість вузлів на шар 250;
- кількість нейронів у прихованих шарах – 250;
- швидкість навчання 0,3;
- коефіцієнт відсіву 0,86;
- кількість повторень – 1500.

Для дослідження оптимізованих гіперпараметрів запропонованої моделі використовується байєсівська оптимізація, оскільки вона

ефективно розкриває загальні максимуми функції чорної скриньки, такі як значення точності для набору перевірок у визначеному просторі параметрів.

Байєсівський підхід відстежує попередні результати оцінки і виводить імовірнісну модель, а потім вибирає наступного кандидата параметрів на основі цієї моделі.

Отже, байєсівська оптимізація дозволяє ефективно шукати оптимальні гіперпараметри. Наше значення цільової функції - точність тестового набору даних. Вхідний шар складається з даних експресії генів. У вихідному рівні є два вузли, оскільки наша проблема полягає у двійковій класифікації, а одне гаряче кодування використовується для вихідної змінної. Вихідний шар містить бінарне значення діагнозу, а саме наявність (1 – «Хворий») чи відсутність (0 – «Здоровий») хвороби Альцгеймера.

ReLU використовується як функція активації, а до вихідного рівня додається шар регресії softmax з оцінками logit для перетворення та нормалізації вихідного значення, яке має бути від 0 до 1.

Модель складається з 8 прихованих шарів з 250 вузлами та одним вузлом зміщення для кожного. Ми використовуємо зменшене середнє значення перехресної ентропії як функцію витрат. Потім ми проводимо оптимізацію градієнтного спуску, щоб мінімізувати витрати.

Коефіцієнти навчання та відсіву в пропонованій моделі встановлюються відповідно 0,02 та 0,85. Максимальна кількість епох – 1500. На рис. 3 можна побачити процес тренування на початкових епохах.

```
DNN parameters
learning_rate: 0.30000
training_epochs: 1500
dr_rate: 0.86
epoch: 0    train_loss: 0.3834  train_acc: 0.8508   test_acc: 0.8235
epoch: 10   train_loss: 0.3782  train_acc: 0.8293   test_acc: 0.8042
epoch: 20   train_loss: 0.3791  train_acc: 0.6651   test_acc: 0.8726
epoch: 30   train_loss: 0.3811  train_acc: 0.8577   test_acc: 0.8263
epoch: 40   train_loss: 0.3815  train_acc: 0.8723   test_acc: 0.7884
```

Рис. 3. Початок тренування нейронної мережі

Щоб уникнути перенавчання, ми застосовуємо не тільки перехресну перевірку, але й ранню зупинку на основі набору тестів під час навчання моделі. Ми визначаємо просте правило для припинення тренувань: через 100 епох для кожної епохи обчислюється середнє значення десяти останніх значень точності тесту, і це значення порівнюється з поточним значенням точності тестування, щоб перевірити, збігається воно чи зменшується. Одночасно, так само, поточна точність навчання порівнюється із середнім значенням тренувань останніх десяти епох, щоб перевірити, чи збільшується вона. Якщо обидва ці правила задоволені, навчання припиняється. На рис. 4 можна побачити взаємозв'язок між втратами тренувань, точністю тренувань та точністю тестування, які були отримані внаслідок тренувань та тестування наборів даних з першого разу.

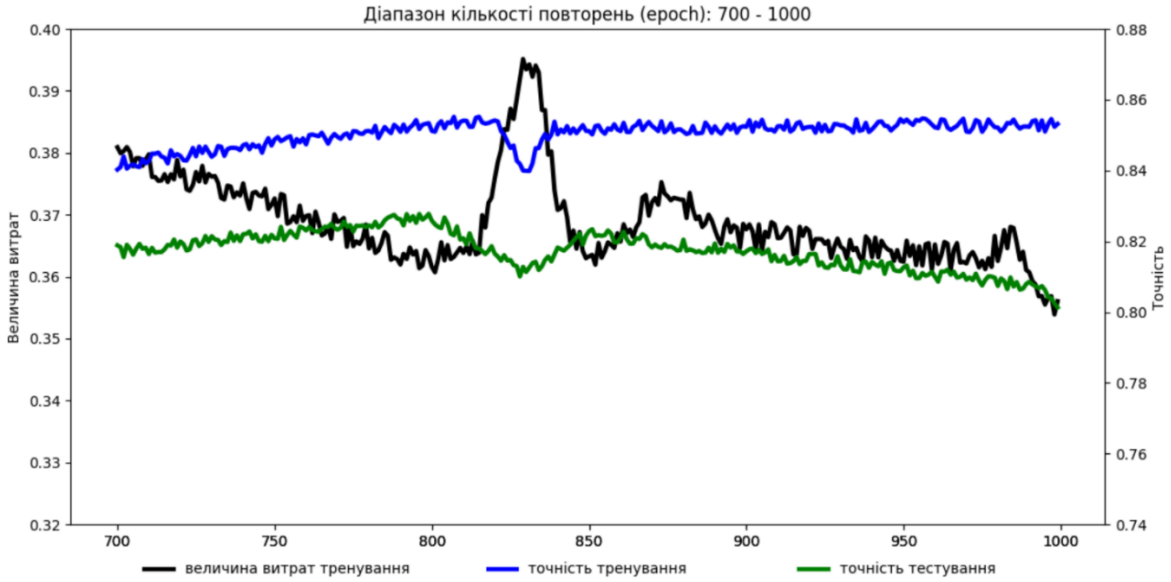

Рис. 4. Величина витрат, точність тренування та тестування в процесі тренування моделі

Як показано на рис., приблизно після 800 епох втрата тренувань зменшилась (чорний), а точність тренувань зросла (синій), при чому точність тестування зменшилась (зелений). Тобто сталося перенавчання.

У нашому підході перенавчання було виявлено в 797-й епосі, і навчання було припинено. Отже, 0,823 була середньою точністю запропонованого глибокого навчання за допомогою нашого методу вибору ознак.

Далі розглянемо складові інтерфейсу користувача програмного додатку. На рис. 5 зображена головна сторінка програми, яка складається з кнопки вибору файлу та прихованої вкладки вимог до структури файлу з даними. Даний інтерфейс дуже легкий для розуміння. Головна сторінка містить назву програмного додатку.

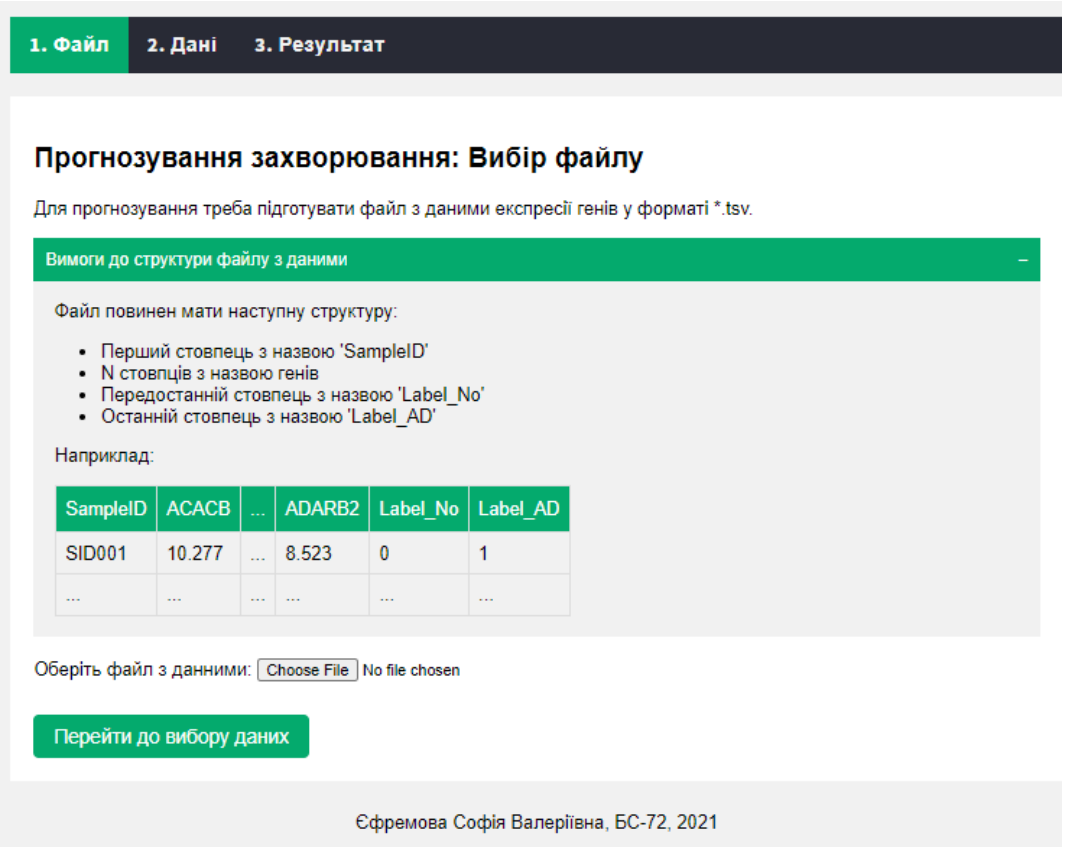

Рис. 5. Приклад інтерфейсу програми

На початку роботі користувач завантажує текстовий файл у форматі *.tsv, в якому знаходяться підготовлені дані експресії генів. Детальніше про вимогу до структури файлу можна побачити при розгортанні вкладки.

На даній вкладці показано вимоги до структури даних для подальшого дослідження, якій надає користувач (науковець). Якщо всі вимоги до задавання файлу виконані правильно слід натиснути кнопку «Перейти до вибору даних» Після чого відкривається наступне вікно програмного додатку, яке зображено на рис. 6.

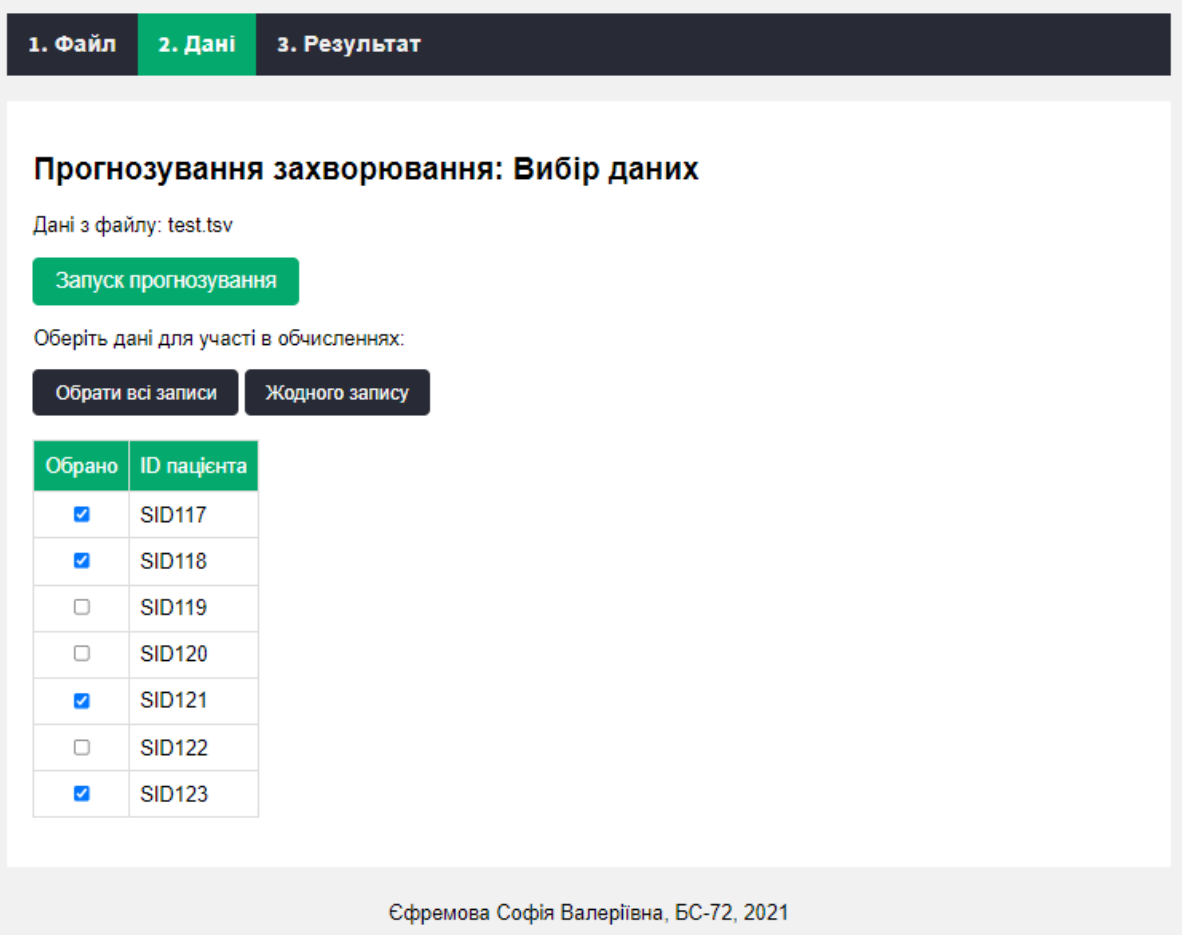

Рис. 6. Повідомлення про порожній файл

На даному вікні дається змога обрати користувачеві обрати кількість пацієнтів для дослідження. Вибір відбувається при установці «галочки» напроти ID потрібного пацієнту. Також можна обрати всіх або жодного пацієнта при натисканні кнопок «Обрати всі записи» та «Жодного запису».

Після обирання ID пацієнтів, які потребують дослідження слід натиснути кнопку «Запуск прогнозування», після чого програма виконує прогнозування та відкриває нове вікно з результатами роботи, які можна побачити на рис. 7.

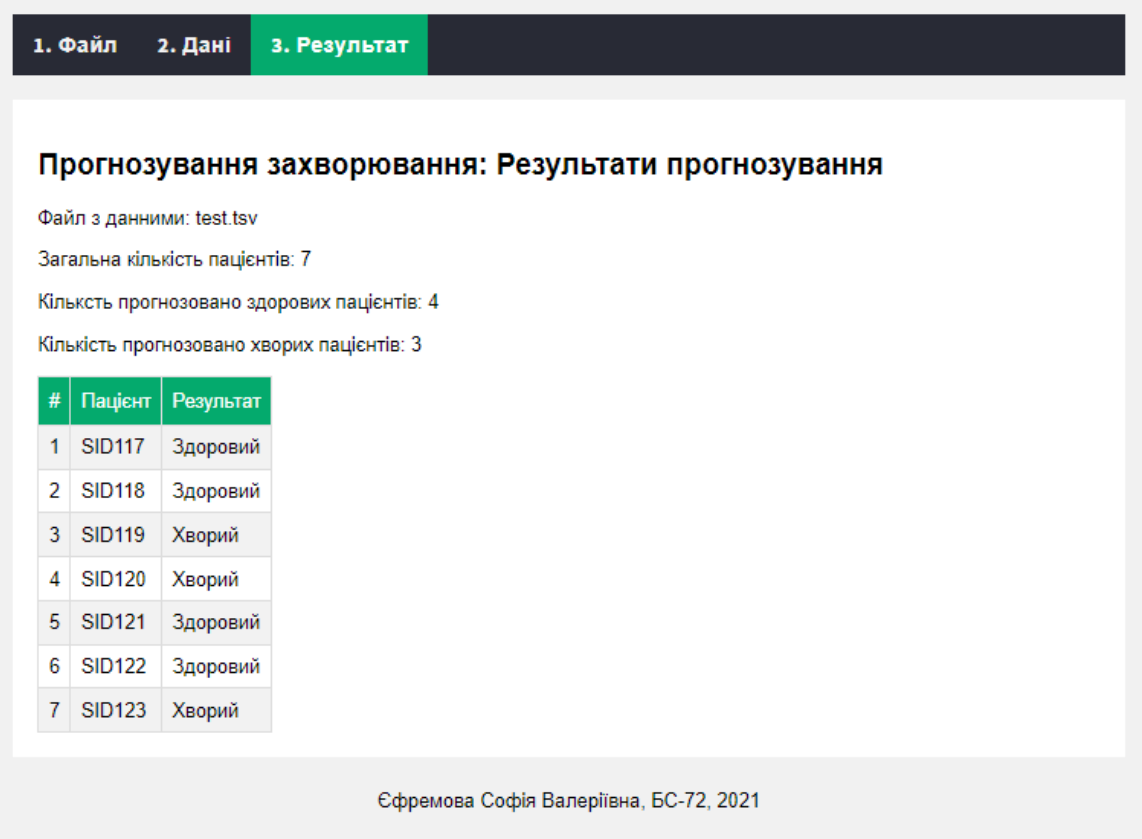

Рис. 7. Результати роботи програмного додатку

Таблиця результатів містить ID пацієнта та результат прогнозування, який має вигляд повідомлення про наявність (Хворий) чи відсутність (Здоровий) хвороби Альцгеймера. Також на цій сторінці виводиться дані про кількість загальну кількість тестованих записів, кількість здорових та хворих на Альцгеймер людей. Це дає змогу користувачеві зрозуміти загальну статистику роботи програми прогнозування хвороби Альцгеймера.

Висновки. В роботі було використано базу NCBI GEO, яка містить дані експресії генів пацієнтів з хворобою Альцгеймера та здорових людей. Початкова база містила 12850 тис. наборів ознак та 180 спостережень. Для зменшення розмірності простору ознак були використані методи дискримінантного аналізу, t-критерію Ст'юдента та кореляційного аналізу. Використання даних методів дозволило скоротити загальну розмірність досліджуваних даних до 250 наборів ознак та 180 спостережень. Для

виконання прогнозування була спроектована глибинна нейронна мережа (DNN). Навчання проводилось 1500 епох, яке дозволило отримати точність 0,823 на екзаменаційній виборці. За допомогою мови програмування Python та мікрофреймворку Flask було створено програмний додаток, який реалізовує можливість прогнозування хвороби Альцгеймера з точністю 0,823.Головним функціоналом програми є індивідуальне чи групове прогнозування хвороби Альцгеймера.

Література

1. Park C., Ha J., Park S. Prediction of Alzheimer's disease based on deep neural network. 2019.
2. Kemle K., Ackermann R.J. Issues in Geriatric Care: Alzheimer Disease. FP Essent. 2018.
3. NCBI GEO. URL: https://www.ncbi.nlm.nih.gov/
4. Osowski S. Methods and tools in data mining. 2013.
5. Sun B.L., Li W.W., Zhu C. та ін.. Clinical Research on Alzheimer's Disease: Progress and Perspectives. Neurosci Bull. 2018.