

УДК 004.8:519.876.5:612.821.1

DOI [https://doi.org/10.24144/2616-7700.2026.48\(1\).172-185](https://doi.org/10.24144/2616-7700.2026.48(1).172-185)**І. М. Лях¹, В. І. Дорогій², М. В. Пономарьов³, Я. О. Чухрай⁴**

¹ ДВНЗ «Ужгородський національний університет»,
професор кафедри інформатики та фізико-математичних дисциплін,
доктор технічних наук, професор
igor.lyah@uzhnu.edu.ua
ORCID: <https://orcid.org/0000-0001-5417-9403>

² ДВНЗ «Ужгородський національний університет»,
аспірант кафедри кібернетики і прикладної математики
viktor.dorohii@uzhnu.edu.ua
ORCID: <https://orcid.org/0009-0005-8174-4822>

³ ДВНЗ «Ужгородський національний університет»,
магістрант 1-ого курсу ОПП «Комп'ютерно-математичне моделювання»
ponomarov.mykhailo@student.uzhnu.edu.ua
ORCID: <https://orcid.org/0009-0005-5064-9232>

⁴ ДВНЗ «Ужгородський національний університет»,
бакалавр 4-ого курсу ОПП «Інформаційні системи та технології»
chukhrai.yaroslav@student.uzhnu.edu.ua
ORCID: <https://orcid.org/0009-0004-6537-1619>

ПРОГНОЗУВАННЯ ПАРАМЕТРІВ СНУ НА ОСНОВІ МАТЕМАТИЧНОГО МОДЕЛЮВАННЯ ТА МАШИННОГО НАВЧАННЯ

У роботі розглядається задача класифікації якості сну за допомогою алгоритмів машинного навчання. Актуальність теми зумовлена зростаючим інтересом до інструментів цифрової медицини, які дозволяють проводити оцінку стану здоров'я людини на основі об'єктивних показників. Метою дослідження є створення ефективної моделі, здатної здійснювати прогноз рівня Quality of Sleep на основі таких факторів, як: вік (Age), тривалість сну (Sleep Duration) та рівень стресу (Stress Level).

На початковому етапі було здійснено комплексну підготовку даних, що включала видалення викидів із використанням методу міжквартильного розмаху, стандартизацію числових змінних, кодування категоріальних ознак та побудову кореляційної матриці. Такий підхід забезпечив підвищення якості навчальних даних та стабільність статистичних оцінок. Далі набір даних було розділено на навчальну та тестову вибірки у співвідношенні 80:20. З метою порівняння ефективності різних підходів до класифікації було реалізовано чотири моделі: Logistic Regression, Random Forest, K-Nearest Neighbors та Support Vector Classifier. Гіперпараметри кожної моделі були підібрані з використанням методу GridSearchCV із п'ятиразовою крос-валідацією.

Оцінювання продуктивності моделей здійснювалося за метриками точності (Accuracy), прецизійності (Precision), повноти (Recall) та F1-міри. Найвищу ефективність показала модель Random Forest, яка досягла значення F1-міри 0.9866. Подальший аналіз включав інтерпретацію ймовірнісних прогнозів моделі та симуляцію нових комбінацій вхідних даних для оцінки впевненості прийнятих рішень. Було встановлено, що більшість прогнозів моделі мають середній або високий рівень ймовірності, що свідчить про її стабільну роботу. Результати дослідження демонструють доцільність використання алгоритмів машинного навчання у завданнях моніторингу якості сну та відкривають перспективи для подальшого вдосконалення моделі шляхом розширення набору ознак і джерел даних.

Ключові слова: машинне навчання, якість сну, метрика, Random Forest, стандартизація.

1. Вступ. У сучасних дослідженнях машинного навчання важливим аспектом є не лише вибір алгоритмів, а й якісна підготовка даних, від якої значною мірою залежить точність та стійкість моделей. Попередня обробка даних охоплює такі етапи, як: видалення викидів, стандартизацію та кодування категоріальних змінних, що забезпечує коректність статистичних оцінок і сприяє підвищенню узагальнюючої здатності моделей. Використання робастного статистичного методу виявлення аномалій, зокрема підходу на основі міжквартильного розмаху, дозволяє усунути екстремальні значення без втрати суттєвої інформації, що підвищує достовірність подальших результатів моделювання.

Після очищення даних виконано стандартизацію для приведення ознак до єдиного масштабу, що особливо важливо при використанні алгоритмів, чутливих до діапазону значень. Категоріальні змінні були перетворені у числовий формат за допомогою кодування міток, що дало змогу забезпечити узгодженість даних із математичними вимогами алгоритмів машинного навчання. Також було побудовано кореляційну матрицю, яка дозволила оцінити взаємозв'язки між змінними та виділити інформативні ознаки для подальшої кластеризації та побудови моделей прогнозування.

Подальші етапи дослідження зосереджено на створенні та оцінюванні моделей класифікації для прогнозування якості сну (Quality of Sleep) на основі ключових факторів – Age (вік), Sleep Duration (Тривалість сну), Stress Level (Рівень стресу). Для цього було застосовано декілька підходів машинного навчання, зокрема логістичну регресію, метод випадкового лісу, k-найближчих сусідів та метод опорних векторів. Їхня ефективність оцінювалася за допомогою стандартних метрик класифікації, що дозволило обґрунтовано визначити найрезультативнішу модель для вирішення поставленої задачі.

2. Постановка завдання. У даному дослідженні поставлено завдання розробити, навчити та порівняти ефективність класифікаційних моделей машинного навчання для прогнозування якості сну (Quality of Sleep) на основі ключових ознак – віку (Age), тривалості сну (Sleep Duration) та рівня стресу (Stress Level). Для досягнення цієї мети було здійснено повний цикл обробки даних: очищення від викидів за допомогою методу міжквартильного розмаху, стандартизацію числових ознак, кодування категоріальних змінних та аналіз кореляційних зв'язків для виявлення релевантних предикторів. На основі підготовлених даних було реалізовано чотири алгоритми класифікації – Logistic Regression, Random Forest, K-Nearest Neighbors та Support Vector Classifier, з підбором гіперпараметрів методом перехресної валідації GridSearchCV. Моделі оцінювалися за метриками Accuracy, Precision, Recall, F1-score, що дозволило визначити Random Forest як найефективніший підхід. Подальший етап передбачав інтерпретацію ймовірнісних прогнозів обраної моделі, виконання зворотного масштабування для відновлення реальних значень ознак та симуляцію прогнозів, що дозволило оцінити рівень впевненості моделі й виявити напрями для її подальшої оптимізації.

3. Аналіз останніх досліджень та публікацій. Вітчизняне дослідження Болобана О. А. присвячене розробці системи прогнозування порушень дихання під час сну з використанням штучного інтелекту. Запропонований підхід поєднує машинне навчання, глибокі нейронні мережі та обробку сигналів фотоплетизмографії для виявлення апное та гіпопное, з адаптивною моделлю LSTM

(Long Short-Term Memory) для аналізу часових рядів. Система на мікросервісній архітектурі підтримує віддалений моніторинг і прийняття діагностичних рішень, відповідаючи сучасним тенденціям ІоМТ (Internet of Medical Things) та цифрової медицини [1].

Дослідження аналізує взаємозв'язок варіабельності серцевого ритму (BCR), стресу та якості сну, показуючи їхній вплив на фізичне й психічне здоров'я. BCR виступає індикатором стану автономної нервової системи, а її зміни можуть слугувати ранніми предикторами серцево-судинних і психоемоційних розладів. Підвищена варіабельність асоціюється з адаптивністю, стресостійкістю та добрим здоров'ям. Проаналізовано механізми впливу стресу на ритмічність серцевої діяльності та якість сну, що поглиблює розуміння патогенетичних зв'язків і окреслює напрями для підтримання психофізіологічного балансу [2].

Wang X. & Wilhelm E. запропонували персоналізовану двоетапну просторово-часову модель прогнозування якості сну, що поєднує конволюційні та рекурентні шари з увагою для захоплення просторових і часових залежностей. Двоетапна адаптація дозволяє підлаштовувати модель під нових користувачів без міток, а порівняння з LSTM та іншими методами показало стабільно кращу точність. Аналіз пояснюваності підтвердив вплив окремих факторів на сон, роблячи підхід персоналізованим і практичним для носимих пристроїв [3].

Дослідження зосереджене на прогнозуванні якості сну за даними носимих пристроїв із використанням глибокого навчання. Запропонована модель WSHMSQP-ODL застосовує Deep Belief Network (DBN) для аналізу даних про активність під час сну, а алгоритм Enhanced Seagull Optimization (ESGO) оптимізує її гіперпараметри. Результати експериментів засвідчили перевагу моделі над аналогами, забезпечуючи точне й персоналізоване прогнозування для систем дистанційного моніторингу та ІоМТ [4].

Дослідження Corda, E., Massa, S. M., & Riboni, D. спрямоване на створення системи прогнозування якості сну з персоналізованими рекомендаціями, що базується на даних сенсорів і алгоритмах машинного навчання (ML). Підхід поєднує ML і великі мовні моделі (LLM), які не лише прогнозують якість сну, а й генерують контекстуальні поради для його покращення [5].

Дослідження Hidayat, A. A., Budiarto, A., Pardamean & B. Long присвячене прогнозуванню якості сну на основі даних про фізичну активність і глибокий сон із носимих пристроїв. Використання LSTM для моделювання часових рядів серцевого ритму та кількості кроків показало, що проста стекована модель дає найточніші результати, тоді як складніші моделі перенавчаються. Це підтверджує ефективність LSTM для прогнозування якості сну за фізіологічними даними [6].

Дослідження Lee, H., Cho, M., Lee, S. W., & Park, S. S. присвячене прогнозуванню якості сну на основі цифрових біомаркерів варіабельності серцевого ритму (HRV), зібраних носимими пристроями. Модель LSTM продемонструвала найвищу точність серед порівнюваних методів, а аналіз LIME виявив ключовий вплив співвідношення LF/HF, показників ISI та WHOQOL-BREF. Робота підтверджує ефективність HRV як біомаркера для персоналізованого моніторингу сну та переваги глибинного навчання у виявленні часових закономірностей у фізіологічних даних [7].

Дослідження Verma, R. K та співавторів аналізує роль штучного інтелекту

у медицині сну, зокрема використання машинного й глибокого навчання для діагностики та управління розладами сну. Автори підкреслюють потенціал ШІ у виявленні апное, інсомнії, нарколепсії та інших порушень, а також його здатність підвищувати доступність діагностики [8].

Дослідження Alabdan R. та співавторів представляє метод MBES-DLSQP для прогнозування якості сну, що поєднує глибоке навчання з модифікованим алгоритмом Bald Eagle Search для оптимізації гіперпараметрів. Модель на основі stacked sparse autoencoder (SSAE) аналізує високорозмірні дані сну з носимих пристроїв і відкритих наборів. Отримані результати підтвердили високу точність прогнозування, демонструючи перспективність підходу для створення зручних і ефективних систем моніторингу сну в медицині [9].

Дослідження Mahnaz Olfati та колег присвячене прогнозуванню тяжкості депресивних симптомів на основі якості сну, рівня тривожності та структурних характеристик мозку. Автори показали, що якість сну є надійним предиктором депресивних проявів, а включення показників тривожності істотно підвищує точність моделей [10].

4. Основний результат. Перед навчанням моделей було проведено огляд та підготовка даних для побудови моделей. Процедура підготовки даних для навчання моделей включали видалення викидів, стандартизацію та кодування категоріальних даних. Для ідентифікації та видалення екстремальних значень, які можуть спотворювати результати моделювання, було використано робастний статистичний метод, що базується на міжквартильному розмахі [11]. Спочатку обчислюється міжквартильний розмах, що є мірою варіації даних:

$$IQR = Q_3 - Q_1, \quad (1)$$

де IQR — міжквартильний розмах; Q_3 — третій квартиль (значення, що відсікає 75% відсортованих даних); Q_1 — перший квартиль (значення, що відсікає 25% відсортованих даних).

На основі IQR визначаються межі, за якими значення вважаються викидами [12]:

$$L = Q_1 - 1.5 \cdot IQR, \quad U = Q_3 + 1.5 \cdot IQR, \quad (2)$$

де L — нижня межа допустимих значень; U — верхня межа допустимих значень.

Усі дані, що виходять за межі діапазону $[L, U]$, відсіюються з набору. Така процедура є непараметричною, тобто не передбачає жодного припущення щодо форми розподілу даних. Вона підвищує стійкість статистичних оцінок (наприклад, середнього та дисперсії) і сприяє покращенню узагальнюючої здатності моделей машинного навчання.

Для приведення даних до єдиного масштабу було виконано стандартизацію. Стандартизація є лінійним перетворенням змінної, спрямованим на приведення даних до єдиного масштабу [13]. Тоді стандартизоване значення визначається як:

$$x'_i = \frac{x_i - \mu}{\sigma}, \quad (3)$$

де x'_i — стандартизоване значення ознаки; x_i — вихідне значення ознаки; μ — середнє арифметичне всіх значень ознаки; σ — стандартне відхилення значень ознаки.

Алгоритми машинного навчання оперують переважно з числовими даними [14]. У зв'язку з цим, категоріальні ознаки, представлені у вигляді текстових міток, потребують попереднього перетворення у числовий формат. Метод кодування міток є одним із фундаментальних підходів для виконання такого перетворення, що полягає у присвоєнні кожній унікальній категорії відповідного унікального цілого числа [15].

Нехай множина унікальних міток змінної позначається як $Y = \{y_1, y_2, \dots, y_m\}$. Процес кодування формально визначається як відображення (функція) f :

$$f : Y \rightarrow \{0, 1, 2, \dots, m - 1\}, \quad (4)$$

що для кожної конкретної мітки y_j виконується за правилом:

$$f(y_j) = j - 1, \quad j = 1, 2, \dots, m, \quad (5)$$

де f — функція кодування, що ставить у відповідність кожній мітці ціле число; $Y = \{y_1, y_2, \dots, y_m\}$ — множина, що складається з m унікальних категоріальних міток; y_j — j -та унікальна мітка з множини Y ; $j - 1$ — цілочисельне значення, що присвоюється мітці y_j , j — порядковий номер мітки (від 1 до m).

Після того як було проведено кодування міток, було побудовано кореляційну матрицю для виявлення та кількісної оцінки лінійних взаємозв'язків між змінними. Це дозволяє зрозуміти, які ознаки змінюються узгоджено (мають спільну динаміку), а які є незалежними, що є важливим етапом перед кластеризацією.

Елементами матриці є коефіцієнти кореляції Пірсона, що розраховується для кожної пари змінних за наступною формулою [16].

$$r_{ij} = \frac{\sum_{k=1}^n (x_{ki} - \bar{x}_i)(x_{kj} - \bar{x}_j)}{\sqrt{\sum_{k=1}^n (x_{ki} - \bar{x}_i)^2} \sqrt{\sum_{k=1}^n (x_{kj} - \bar{x}_j)^2}}, \quad (6)$$

де r_{ij} — коефіцієнт кореляції Пірсона між змінними x_i та x_j ; n — кількість спостережень у вибірці; x_{ki} — k -те спостереження змінної x_i ; x_{kj} — k -те спостереження змінної x_j ; \bar{x}_i та \bar{x}_j — середні арифметичні значення змінних x_i та x_j відповідно.

Значення коефіцієнтів $r_{ij} \in [-1, 1]$, де 1 означає ідеальний прямий лінійний зв'язок, -1 — ідеальний обернений, а 0 — відсутність лінійного зв'язку. Кореляційна матриця необхідна для відбору ознак перед побудовою кластеризації.

Для візуального представлення кореляційної матриці використовується теплокорта. Це графічний інструмент, де кожному значенню коефіцієнта кореляції відповідає певний колір або відтінок. Такий підхід дозволяє швидко та інтуїтивно виявляти сильні позитивні (зазвичай позначаються теплими кольорами) та негативні (холодними кольорами) зв'язки, а також ідентифікувати групи взаємопов'язаних ознак.

Результати проведеного кореляційного аналізу для досліджуваного набору даних представлено на рис. 1.

Ключовим етапом у побудові будь-якої моделі машинного навчання є визначення цільової змінної (залежної змінної) та набору ознак (незалежні змінні), які будуть використовуватись для її прогнозування.

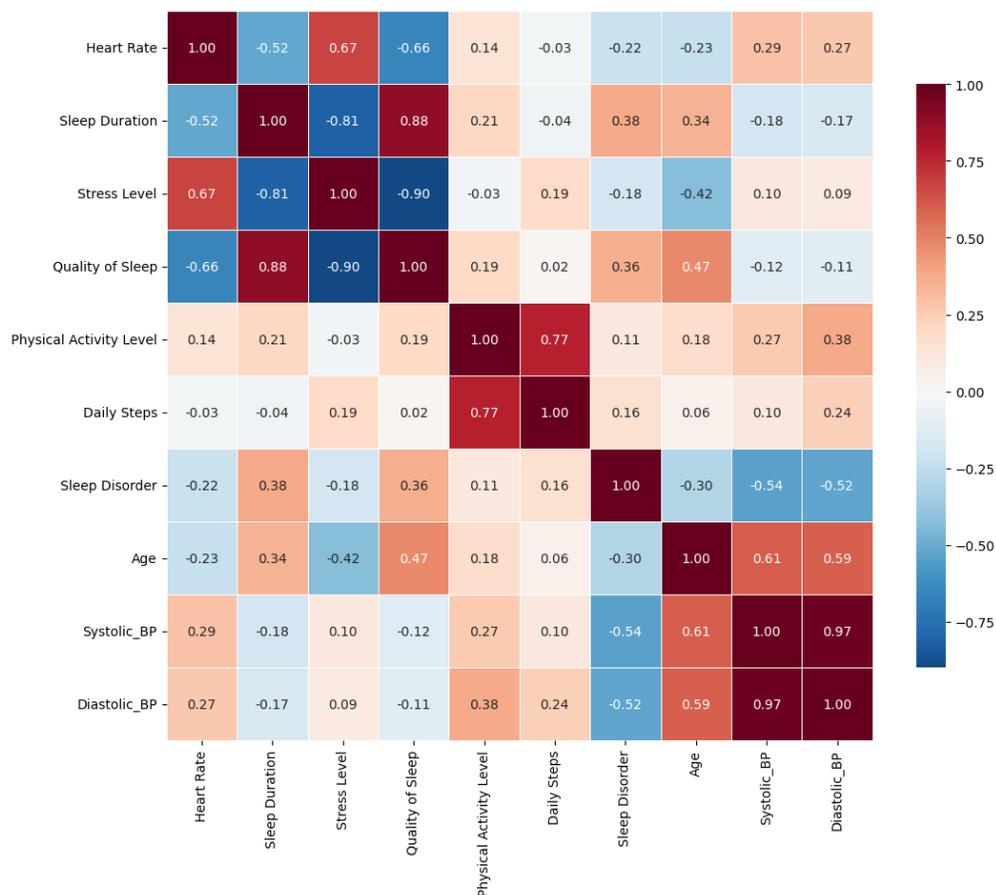


Рис. 1. Кореляційна матриця.

У даному дослідженні в якості цільової змінної було обрано показник *Quality of Sleep*. Ця змінна є дискретною оцінкою якості сну, і саме її модель буде намагатися спрогнозувати на основі інших факторів. В якості незалежних змінних було обрано наступні ознаки: *Age* (вік), *Sleep Duration* (Тривалість сну), *Stress Level* (Рівень стресу).

Після визначення цільової змінної та набору ключових ознак вибірку поділено на дві незалежні частини: навчальну та тестову.

Формально, якщо вихідний набір даних позначити як D , то його розбиття відповідає наступним умовам:

$$D = D_{train} \cup D_{test}, \quad D_{train} \cap D_{test} = \emptyset, \quad (7)$$

де D — повний набір даних; D_{train} — навчальна вибірка (у нашому випадку 80% даних); D_{test} — тестова вибірка (відповідно, 20% даних);

Після розбиття даних, для досягнення максимальної ефективності моделей необхідно було підібрати їхні оптимальні гіперпараметри — налаштування, що задаються перед навчанням. Для систематичного пошуку найкращої конфігурації було застосовано метод пошуку по сітці (GridSearchCV) з перехресною валідацією [17].

Цей процес полягає у пошуку набору гіперпараметрів θ^* , що мінімізує усереднену похибку, оцінену методом k -fold перехресної валідації [18]. Для цього

навчальна вибірка ділиться на k блоків (у випадку дослідження $k = 5$), і для кожної комбінації параметрів модель послідовно навчається на $k - 1$ блоках з валідацією на залишковому. Усереднена похибка по всіх ітераціях $CV(\theta)$ слугує стійкою оцінкою якості:

$$CV(\theta) = \frac{1}{k} \sum_{i=1}^k L_i(\theta), \quad (8)$$

де $CV(\theta)$ — усереднена похибка перехресної валідації; k — кількість блоків; $L_i(\theta)$ — похибка моделі, обчислена на i -му блоці валідації.

Метод пошуку по сітці (GridSearchCV) автоматизує цей процес, послідовно перебираючи всі задані комбінації гіперпараметрів та обираючи той, який покаже найкращий результат $CV(\theta)$. Саме ця оптимальна конфігурація використовувалася для навчання кінцевої моделі.

Після визначення загальної стратегії підбору гіперпараметрів, необхідно детально розглянути архітектуру та математичні принципи кожного з алгоритмів класифікації, що були застосовані в рамках даного дослідження для порівняльного аналізу їхньої ефективності.

У дослідженні було використано чотири різні алгоритми, кожен з яких має унікальні математичні підходи до вирішення задачі класифікації.

Логістична регресія (Logistic Regression) — це статистичний метод, що моделює ймовірність належності об'єкта до певного класу за допомогою логістичної функції (сигмоїди) [19].

$$P(x) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)}}, \quad (9)$$

де $P(x)$ — ймовірність належності об'єкта до позитивного класу; β_i — вагові коефіцієнти.

Випадковий ліс (Random Forest) — це ансамблевий метод, що усереднює прогнози множини незалежних дерев рішень, кожне з яких побудоване на випадковій підмножині даних [20]. Підсумкове рішення приймається шляхом голосування:

$$\hat{y} = \text{mode}\{h_1(x), h_2(x), \dots, h_T(x)\}, \quad (10)$$

де \hat{y} — фінальний прогноз $h_t(x)$ — передбачення t -го дерева; T — загальна кількість дерев.

Метод k -найближчих сусідів (k -NN) класифікує об'єкт на основі класу більшості з його k найближчих сусідів у просторі ознак [21]. Відстань між об'єктами зазвичай обчислюється за евклідовою метрикою:

$$d(x_i, x_j) = \sqrt{\sum_{k=1}^n (x_{ik} - x_{jk})^2}, \quad (11)$$

де $d(x_i, x_j)$ — евклідова відстань між об'єктами x_i та x_j ; n — кількість ознак.

Метод опорних векторів (SVC) шукає оптимальну гіперплощину, що максимізує відстань (margin) між класами [22]:

$$\max_{w,b} \frac{2}{|w|} \text{ за умови } y_i (w^T x_i + b) \geq 1, \quad (12)$$

де w — вектор нормалі до гіперплощини; b — зсув гіперплощини; $w \vee$ — норма вектора w ; x_i та y_i — вектор ознак та мітка класу i -го об'єкта. Для нелінійних задач застосовується ядрове перетворення $K(x_i, x_j)$, яке дозволяє знайти розділення у вищому вимірі.

Після того, як кожна модель була навчена з оптимальними гіперпараметрами, наступним етапом є об'єктивна оцінка та порівняння їхньої продуктивності на тестовій вибірці. Для цього використовується набір стандартних метрик класифікації, таких як: точність (Accuracy), повнота (Recall), прецизійність (Precision), F1-міра (F1-Score).

Основою для їх розрахунку слугує матриця невідповідностей (confusion matrix), яка деталізує результати класифікації за чотирма категоріями: TP (True Positive) — істинно позитивні випадки; TN (True Negative) — істинно негативні; FP (False Positive) — хибно позитивні; FN (False Negative) — хибно негативні.

На основі цих компонентів матриці розраховується ряд ключових метрик. Точність (Accuracy) відображає загальну частку коректно класифікованих об'єктів відносно загальної кількості спостережень.

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}. \quad (13)$$

Для більш глибокого аналізу, особливо в умовах незбалансованих класів, використовується інша метрика, а саме прецизійність (Precision), яка оцінює надійність позитивних прогнозів.

$$\text{Precision} = \frac{TP}{TP + FP}. \quad (14)$$

У свою чергу, метрика повнота (Recall або Чутливість) вимірює здатність моделі ідентифікувати всі фактично позитивні об'єкти з вибірки.

$$\text{Recall} = \frac{TP}{TP + FN}. \quad (15)$$

Оскільки між прецизійністю та повнотою часто існує обернений зв'язок, для їх узагальнення використовується F1-міра (F1-Score). Ця метрика є гармонійним середнім між ними і дозволяє знайти оптимальний баланс.

$$F1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (16)$$

Для наочної демонстрації, на рис. 2 представлено матрицю невідповідностей для розробленої моделі Random Forest.

Для кожної з чотирьох розроблених моделей були розраховані вищезазначені метрики на тестовій вибірці. Підсумкові результати оцінки всіх моделей представлено в табл. 1.



Рис. 2. Матриця невідповідностей для Random Forest.

Таблиця 1.

Порівняльна таблиця метрик якості для класифікаційних моделей

Model	Accuracy	Precision	Recall	F1-score
Logistic Regression	0.866667	0.843095	0.866667	0.853581
Random Forest	0.986667	0.987246	0.986667	0.9866
KNN	0.946667	0.938739	0.946667	0.939363
SVC	0.933333	0.928811	0.933333	0.925234

Комплексний аналіз результатів, представлених у таблиці, дозволив визначити Random Forest як найбільш ефективну модель для даної задачі. Він не тільки показав високу загальну точність, але й досяг найкращого балансу між Precision та Recall, що відображається у найвищому значенні F1-міри.

Після вибору найкращої моделі – Random Forest, кінцевий етап дослідження полягає в інтерпретації цих результатів та демонстрації практичного застосування. Це включає повернення даних до початкового масштабу, аналіз імовірнісних оцінок та формування конкретних класифікаційних прогнозів.

На етапі попередньої обробки даних було застосовано стандартизацію для приведення всіх ознак до єдиного масштабу за формулою (3). Для того, щоб результати моделювання були інтерпретованими у своїх вихідних одиницях, виконується процедура зворотного масштабування.

Дана операція є математично оберненою до стандартизації та дозволяє відновити початкове значення ознаки.

$$x_i = x'_i \cdot \sigma + \mu, \quad (17)$$

де x_i — відновлене значення у початковому масштабі; x'_i — стандартизоване значення; σ — стандартне відхилення, що використовувалось на етапі стандартизації; μ — середнє арифметичне, що використовувалось на етапі стандартизації.

Обрана модель Random Forest використовується для виконання класифікаційного прогнозу, тобто для визначення найбільш імовірного класу *Quality of Sleep* для набору вхідних даних (*Age, Sleep Duration, Stress Level*). У процесі класифікації, окрім визначення передбаченого класу, модель форм у ймовірнісну оцінку належності об'єкта до кожної категорії. Це числове значення в межах від 0 до 1 відображає ступінь упевненості моделі у своєму прогнозі та інтерпретується як апостеріорна ймовірність належності об'єкта до певного класу. Висока ймовірність свідчить про високий рівень достовірності прогнозу, тоді як низька вказує на невизначеність або можливу належність об'єкта до іншого класу.

Для інтерпретації рівня впевненості доцільно використовувати шкалу ймовірностей:

0.00 – 0.40 — низька (модель не впевнена у прогнозі);

0.41 – 0.60 — середня (потрібна додаткова перевірка);

0.61 – 0.80 — висока (модель здебільшого впевнена);

0.81 – 1.00 — дуже висока (прогноз майже однозначний).

Для демонстрації практичного застосування моделі було проведено симуляцію прогнозування якості сну, в рамках якої генерувалися різні комбінації вхідних ознак. Для кожного набору даних модель надавала класифікаційний прогноз та відповідну йому ймовірність. Отримані результати дали змогу оцінити ступінь надійності передбачень та визначити, у яких випадках модель демонструє найвищу впевненість. Дані результати представлені в табл. 2.

Таблиця 2.

Приклад прогнозів якості сну з відповідними ймовірнісними оцінками

Age	Sleep Duration	Stress Level	Predicted Class	Probability
20	5.1	7	4	0.46
21	5.9	7	4	0.45
34	6.8	7	4	0.49
29	6.7	7	5	0.72
30	8.8	7	5	0.33
36	6.7	7	5	0.74
39	5.5	6	6	0.54
31	5.1	6	6	0.59
43	5.5	7	6	0.7
30	9.3	6	7	0.73
26	9.2	6	7	0.64
47	8.5	7	7	0.46
41	8.4	5	8	0.5
42	7.5	3	8	0.54
33	9.1	5	8	0.425
42	8.1	4	9	0.62
47	9.4	2	9	0.86
62	9.3	5	9	0.9

Дана табл. 2 наочно ілюструє, як модель реагує на різні вхідні дані, та дозволяє оцінити не тільки прогнозований результат, але й ступінь впевненості моделі, що є критично важливим для прийняття рішень у практичних задачах.

Згідно з отриманими даними, найбільша частка значень (50%) припадає на інтервал 0.41 – 0.60, який характеризує середній рівень упевненості. Це свідчить про те, що в більшості випадків модель формує прогнози з помірною достовірністю, що може вказувати на наявність частково неоднозначних вхідних даних. Високий рівень упевненості (0.61 – 0.80) становить 33.3%, а дуже високий (0.81 – 1.00) – лише 11%, що свідчить про обмежену кількість ситуацій, у яких модель повністю впевнена у своєму рішенні. Низькі значення ймовірностей (0.00 – 0.40) трапляються рідко – менше 6% від загальної кількості випадків, що свідчить про відносну стабільність алгоритму.

Отже, аналіз розподілу ймовірностей підтверджує, що модель демонструє збалансовану поведінку, проте потребує подальшого вдосконалення з метою підвищення частки прогнозів із високим рівнем впевненості.

5. Висновки та перспективи подальших досліджень. У межах дослідження було реалізовано повний цикл побудови системи класифікації якості сну на основі даних про вік, тривалість сну та рівень стресу. Проведено ґрунтовну передобробку даних, яка включала видалення викидів методом міжквартильного розмаху, стандартизацію числових ознак, кодування категоріальних змінних та аналіз кореляційних зв'язків між змінними. Це забезпечило належну якість вхідних даних і підвищило достовірність подальшого моделювання. Було сформовано навчальну та тестову вибірки, на основі яких здійснено навчання чотирьох класифікаційних моделей: Logistic Regression, Random Forest, K-Nearest Neighbors і Support Vector Classifier з оптимізацією гіперпараметрів за допомогою GridSearchCV.

Порівняльна оцінка моделей за метриками точності, прецизійності, повноти та F1-міри засвідчила, що модель Random Forest продемонструвала найвищу ефективність на тестовій вибірці, досягнувши F1-міри 0.9866. Це свідчить про здатність моделі забезпечувати не лише високу загальну точність, але й збалансованість між кількістю правильно виявлених позитивних класів та надійністю класифікаційних рішень. Подальша симуляція прогнозів і аналіз ймовірнісних оцінок дозволили дослідити ступінь впевненості моделі у своїх передбаченнях, що є важливою умовою практичного застосування в умовах неповної або неоднозначної інформації.

Аналіз розподілу ймовірностей показав, що більшість прогнозів моделі мають середній або високий рівень упевненості, а частка випадків із дуже високою впевненістю становить понад 10%. Це свідчить про стабільну роботу моделі, однак також вказує на можливість подальшого підвищення точності через розширення ознакового простору або застосування методів балансування класів. Отримані результати демонструють потенціал використання алгоритмів машинного навчання для прогнозування стану сну та можуть бути основою для впровадження персоналізованих рекомендацій у сфері цифрової медицини та добробуту.

Конфлікт інтересів

Автори заявляють, що не мають конфлікту інтересів щодо даного дослідження, включаючи фінансовий, особистий, авторський або будь-який інший, який міг би вплинути на дослідження, а також на результати, представлені в даній статті.

Фінансування

Дослідження було проведено без фінансової підтримки.

Доступність даних

Усі дані доступні в цифровій або графічній формі в основному тексті рукопису.

Використання штучного інтелекту

Автори підтверджують, що при створенні даної роботи вони не використовували технології штучного інтелекту.

Внесок авторів

І. М. Лях – концептуалізація, підготовка даних, формальний аналіз, написання – оригінальний проєкт, науковий нагляд; В. І. Дорогий – методологія, адміністрування проєкту, написання (рецензування та редагування); М. В. Пономарьов – проведення дослідження, ресурси, візуалізація; Я. О. Чухрай – валідація, програмне забезпечення.

Авторські права ©



(2026). Лях І. М., Дорогий В. І., Пономарьов М. В., Чухрай Я. О. Ця робота ліцензується відповідно до Creative Commons Attribution 4.0 International License.

Список використаної літератури

1. Boloban, O. A. (2025). *Methods and tools for predicting respiratory diseases and apnea based on artificial intelligence and microservice architecture*: (dissertation ...doctor of philosophy). Kyiv, 156 p. Retrieved from <https://ela.kpi.ua/items/7d1c9b55-5f5b-4bd2-97cc-32341abe501f> [in Ukrainian].
2. Liashenko, V. P., & Stetsenko, S. M. (2024). Peculiarities of heart rate variability against the background of sleep disturbance and stress factors: theoretical aspect. *Slobozhansky Scientific Bulletin. Series: Natural Sciences*, (1), 43–49. <https://doi.org/10.32782/naturalspu/2024.1.5>
3. Wang, X., & Wilhelm, E. (2025). Individualized and Interpretable Sleep Forecasting via a Two-Stage Adaptive Spatial-Temporal Model. *arXiv preprint*.

- <https://doi.org/10.48550/arXiv.2509.06974>
4. Hamza, M. A., Abdalla Hashim, A. H., Alsolai, H., Gaddah, A., Othman, M., Yaseen, I., ... & Zamani, A. S. (2023). Wearables-assisted smart health monitoring for sleep quality prediction using optimal deep learning. *Sustainability*, 15(2), 1084. <https://doi.org/10.3390/su15021084>
 5. Corda, E., Massa, S. M., & Riboni, D. (2024). Context-aware behavioral tips to improve sleep quality via machine learning and large language models. *Future Internet*, 16(2), 46. <https://doi.org/10.3390/fi16020046>
 6. Hidayat, A. A., Budiarto, A., & Pardamean, B. (2023). Long short-term memory-based models for sleep quality prediction from wearable device time series data. *Procedia Computer Science*, 227, 1062–1069. <https://doi.org/10.1016/j.procs.2023.10.616>
 7. Lee, H., Cho, M., Lee, S. W., & Park, S. S. (2025). Predicting sleep quality with digital biomarkers and artificial neural networks. *Frontiers in Psychiatry*, 16, 1591448. <https://doi.org/10.3389/fpsy.2025.1591448>
 8. Verma, R. K., Dhillon, G., Grewal, H., Prasad, V., Munjal, R. S., Sharma, P., ... & Surani, S. (2023). Artificial intelligence in sleep medicine: Present and future. *World Journal of Clinical Cases*, 11(34), 8106. <https://doi.org/10.12998/wjcc.v11.i34.8106>
 9. Alabdan, R., Mengash, H. A., Maray, M., Alotaibi, F., Abdelbagi, S., & Mahmud, A. (2023). Modified bald eagle search algorithm with deep learning-driven sleep quality prediction for healthcare monitoring systems. *IEEE Access*, 11, 135385–135393. <https://doi.org/10.1109/ACCESS.2023.3337647>
 10. Olfati, M., Samea, F., Faghihroohi, S., Balajoo, S. M., Küppers, V., Genon, S., ... & Tahmasian, M. (2024). Prediction of depressive symptoms severity based on sleep quality, anxiety, and gray matter volume: a generalizable machine learning approach across three datasets. *EBioMedicine*, 108. <https://doi.org/10.1016/j.ebiom.2024.105313>
 11. Loh, P. L. (2024). A theoretical review of modern robust statistics. *Annual Review of Statistics and Its Application*, 12. <https://doi.org/10.1146/annurev-statistics-112723-034446>
 12. Ding, Y. J., Yap, W. S., & Khor, K. C. (2023). Profiling and identifying Smurfs or boosters on Dota 2 using K-means and IQR. *IEEE Transactions on Games*, 16(3), 577–585. <https://doi.org/10.1109/TG.2023.3317053>
 13. Divakaruni, A. S., & Jastroch, M. (2022). A practical guide for the analysis, standardization and interpretation of oxygen consumption measurements. *Nature metabolism*, 4(8), 978–994. <https://doi.org/10.1038/s42255-022-00619-4>
 14. Kavetskyi, M. S., Sievierinov, O. V., Gvozdo, R. Y., & Smirnov, A. O. (2024). Using machine learning to classify DOS/DDOS attacks. *Radiotekhnika*, 2(217), 55–63. <https://doi.org/10.30837/rt.2024.2.217.04> [in Ukrainian].
 15. Mishchenko, L. D., & Klymenko, I. A. (2023). Recognition of fake news using natural language processing and a low-power architecture for edge computing. *Problems of Informatization and Management*, 4(76), 59–67. <https://doi.org/10.18372/2073-4751.76.18241> [in Ukrainian].
 16. Kotenko, V. V., Bashynskiy, S. I., & Piskun, I. A. (2021). Application of the Pearson method to obtain dependences of the chemical elements distribution within the kaolin deposit. *Technical Engineering*, (2(88)), 129–134. [https://doi.org/10.26642/ten-2021-2\(88\)-129-134](https://doi.org/10.26642/ten-2021-2(88)-129-134) [in Ukrainian].
 17. Aprilliandhika, W., & Abdulloh, F. F. (2024). Comparison Of K-Nearest Neighbor and Support Vector Machine Algorithm Optimization With Grid Search CV On Stroke Prediction. *Jurnal Teknik Informatika (Jutif)*, 5(4), 991–1000. <https://doi.org/10.52436/1.jutif.2024.5.4.1951>
 18. Oursatye, O. A., Volkov, O. Y., & Tkalya, V. H. (2025). Automated Machine Learning. State and Prospects Development. *Information Technologies and Systems*, 2(2), 3–33. <https://doi.org/10.15407/intechsys.2025.02.003> [in Ukrainian].
 19. Radionova, I., & Fedorenko, T. (2025). Analysis of state budget expenditures in wartime using logistic regression. *Science Notes of KROK University*, (3(79)), 25–34. <https://doi.org/10.31732/2663-2209-2025-79-25-34> [in Ukrainian].
 20. Liakh, I. M., Dudnyk, V. V., Tsipino, Y. M., & Tsipino, A. Y. (2025). Forecasting user engagement in educational web platforms using machine learning algorithms. *Scientific*

- Bulletin of Uzhhorod University. Series of Mathematics and Informatics*, 46(1), 218–225. [https://doi.org/10.24144/2616-7700.2025.46\(1\).218-225](https://doi.org/10.24144/2616-7700.2025.46(1).218-225) [in Ukrainian].
21. Klymenko, O. (2024). Concept of supervised machine learning algorithms based on the k-nearest neighbors method for load balancing on pharmaceutical sorting lines. *Control, Navigation and Communication Systems*, 1(75), 28–30. <https://doi.org/10.26906/SUNZ.2024.1.028> [in Ukrainian].
 22. Bovchaliuk, S., & Haidai, Y. (2024). Analysis of the support vector machine algorithm in comparison to traditional market movements prediction methods. *Control, Navigation and Communication Systems*, 3(77), 89–92. <https://doi.org/10.26906/SUNZ.2024.3.089> [in Ukrainian].

Liakh I. M., Dorohii V. I., Ponomarov M. V., Chukhrai Y. O. Forecasting Sleep Parameters Based on Mathematical Modelling and Machine Learning.

This study addresses the task of classifying sleep quality using machine learning algorithms. The relevance of the topic arises from the growing interest in digital medicine tools that enable the assessment of human health based on objective indicators. The aim of the research is to develop an efficient model capable of predicting the level of «Quality of Sleep» based on factors such as age, sleep duration, and stress level.

At the initial stage, a comprehensive data preprocessing procedure was conducted, which included the removal of outliers using the interquartile range method, standardisation of numerical variables, encoding of categorical features, and the construction of a correlation matrix. This approach ensured improved data quality and enhanced the stability of statistical estimates. Subsequently, the dataset was divided into training and testing subsets in an 80:20 ratio. To compare the effectiveness of different classification approaches, four models were implemented: Logistic Regression, Random Forest, K-Nearest Neighbours, and Support Vector Classifier. The hyperparameters of each model were optimised using the GridSearchCV method with five-fold cross-validation.

The evaluation of the model's performance was carried out using accuracy, precision, recall, and F1-score metrics. The Random Forest model demonstrated the highest effectiveness, achieving an F1-score of 0.9866. Further analysis included the interpretation of the model's probabilistic predictions and the simulation of new combinations of input data to assess the confidence of the generated decisions. It was found that most of the model's predictions exhibited medium or high probability levels, indicating its stable performance. The results of the study demonstrate the feasibility of applying machine learning algorithms to sleep quality monitoring tasks and highlight the potential for further model enhancement through the expansion of feature sets and data sources.

Keywords: machine learning, quality of sleep, metrics, Random Forest, standardisation.

Отримано: 15.08.2025

Прийнято: 13.11.2025

Опубліковано: 29.01.2026