НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Факультет прикладної математики

Кафедра прикладної математики

«На правах рукопису»

УДК 519.2:616.379-008.64

«До захисту допущено»

Завідувач кафедри

_____ О. Р. Чертов

«___» _____ 2016 р.

Магістерська дисертація

на здобуття ступеня магістра

зі спеціальності 8.04030101 «Прикладна математика»

на тему: Прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу

Виконали: студенти II курсу, групи КМ-41м

Сахаров Сергій Юрійович _____

Юрченко Дмитро Володимирович _____

Науковий керівник зав. кафедри, д-р техн. наук, доцент _____
Чертов О. Р.

Консультант із нормоконтролю старший викладач Мальчиков В. В. _____

Рецензент професор, д-р техн. наук, проф. _____
Кулаков Ю. О.

Засвідчую, що в цій магістерській дисертації немає запозичень із праць інших авторів без відповідних посилань.
Сахаров С. Ю. _____
Юрченко Д. В. _____

Київ — 2016

Національний технічний університет України

«Київський політехнічний інститут»

Факультет прикладної математики

Кафедра прикладної математики

Рівень вищої освіти — другий (магістерський)

Спеціальність 8.04030101 «Прикладна математика»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ О. Р. Чертов

«___» _____ 2016 р.

ЗАВДАННЯ

на магістерську дисертацію студентам

Юрченку Дмитру Володимировичу

Сахарову Сергію Юрійовичу

1. Тема дисертації: «Прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу», науковий керівник дисертації Чертов Олег Романович, д-р техн. наук, доцент, затверджені наказом по університету від «21» березня 2016 р. № 1187-С.

2. Термін подання студентом дисертації: «10» червня 2016 р.

3. Об'єкт дослідження: математичні методи прогнозування нічної гіпоглікемія у хворих на діабет 1-го типу.

4. Предмет дослідження: розробка та дослідження математичного методу прогнозування нічних приступів гіпоглікемії у хворих на цукровий діабет 1-го типу на основі фізіологічних і демографічних показників за допомогою методів машинного навчання.

5. Перелік завдань, які потрібно розробити: розробити методику виділення ключових значень з часового ряду показів CGM, проаналізувати існуючі математичні методи прогнозування, систематизувати існуючі математичні методи прогнозування нічної гіпоглікемії у хворих на діабет першого типу, обрати та пристосувати методи на основі машинного навчання для вирішення задачі прогнозування нічної гіпоглікемії,

створити програмне забезпечення, що реалізує обрані методи машинного навчання, провести експериментальне дослідження створеного програмного забезпечення на клінічних даних хворих на діабет першого типу.

6. Орієнтовний перелік ілюстративного матеріалу: існуючі методи прогнозування нічної гіпоглікемії, методи вирішення задачі прогнозування, приклад виділення ключових значень відповідно до розробленої методики виділення ключових значень з часового ряду показів CGM, таблиця часового ряду показіва CGM проекту DirecNet, таблиця, отримана в результаті відбору значень з часового ряду, приклад матриці похибок.

7. Орієнтовний перелік публікацій: тези «Виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу», тези «Прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу на основі дерев прийняття рішень», тези «Методика прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу за допомогою нейронних мереж».

8. Дата видачі завдання: «2» березня 2016 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Ґрунтовне ознайомлення з предметною областю	15.12.2014	
2	Визначення структури магістерської дисертації; вивчення літератури, пошук додаткової літератури	01.03.2015	
3	Робота над першим, другим та третім розділами спільної частини магістерської дисертації	15.05.2015	
4	Проведення наукового дослідження; робота над четвертим розділом спільної частини магістерської дисертації	15.10.2015	

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
5	Проведення наукового дослідження; робота над статтею за результатами наукового дослідження	15.12.2015	
6	Робота над першим, другим, третім та четвертим розділами індивідуальних частин магістерської дисертації; підготовка статті за результатами наукового дослідження; розроблення програмного забезпечення	01.03.2016	
7	Завершення роботи над основною частиною магістерської дисертації	15.05.2016	
8	Оформлення текстової і графічної частин магістерської дисертації	25.05.2016	

Студент _____ Сахаров С. Ю.

Студент _____ Юрченко Д. В.

Науковий керівник дисертації _____ Чертов О. Р.

РЕФЕРАТ

Дисертацію виконано на 73 аркушах, вона містить 2 додатки та перелік посилань на використані джерела з 39 найменувань. У роботі наведено 22 рисунки та 9 таблиць.

Актуальність теми. Цукровий діабет — це хронічне захворювання, яке потребує постійного медичного догляду і нагляду з боку самого хворого, щоб попередити можливі ускладнення та зменшити ризик довгострокових ускладнень. Згідно з даними International Diabetes Federation (IDF), в світі нараховується більше 415 мільйонів хворих на діабет людей. Гіпоглікемія є нагальною проблемою для хворих на діабет першого типу (тобто таких, організм котрих не може самостійно виробляти інсуліну). Відповідно до статистики, хворі на діабет першого типу мають в середньому два приступи симптоматичної гіпоглікемії кожного тижня і один тяжкий приступ гіпоглікемії один раз на рік.

Прогнозування приступів нічної гіпоглікемії у хворих є необхідним для попередження падіння рівня глюкози в плазмі крові нижче норми у нічний час доби. У випадку падіння рівня глюкози нижче норми, функціонування організму порушується, що може призвести до смерті. Тому створення методів прогнозування нічної гіпоглікемії є важливою задачею, в результаті вирішення якої можна зменшити ризики для життя хворих на цукровий діабет.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась згідно з планом науково-дослідних робіт кафедри прикладної математики Національного технічного університету України «Київський політехнічний інститут».

Мета і задачі дослідження. Метою дисертаційної роботи є розробка математичних методів прогнозування нічної гіпоглікемії для попередження приступів нічної гіпоглікемії у хворих на діабет першого типу.

Для досягнення вказаної мети було розв'язано такі задачі:

– розробити методику виділення ключових значень з часового ряду показів CGM;

– проаналізувати існуючі математичні методи прогнозування;

– систематизувати існуючі математичні методи прогнозування нічної гіпоглікемії у хворих на діабет першого типу;

– обрати та пристосувати методи на основі машинного навчання для вирішення задачі прогнозування нічної гіпоглікемії;

– створити програмне забезпечення, що реалізує обрані методи машинного навчання;

– провести експериментальне дослідження створеного програмного забезпечення на клінічних даних хворих на діабет першого типу.

Об'єктом дослідження є математичні методи прогнозування нічної гіпоглікемія у хворих на діабет 1-го типу.

Предметом дослідження є розробка та дослідження математичного методу прогнозування нічних приступів гіпоглікемії у хворих на цукровий діабет 1-го типу на основі фізіологічних і демографічних показників за допомогою методів машинного навчання.

Методи дослідження. Для розв'язання поставленої задачі використовувалися такі методи: методи машинного навчання (для розроблення методів розв'язання задачі прогнозування нічної гіпоглікемії у хворих на діабет першого типу); методи теорії алгоритмів та програмування (для програмної реалізації розроблених алгоритмів); методи теорії ймовірності та математичної статистики (для аналізу результатів експериментів).

Наукова новизна одержаних результатів складається з наступних положень:

– уперше використано демографічні дані та час замірів рівня глюкози в крові разом зі значеннями рівня глюкози в крові для побудови прогнозу, на відміну від існуючих методів, де використовується лише значення рівня глюкози в крові;

– уперше розроблено методику виокремлення ключових значень рівня глюкози в крові з показань пристроїв CGM;

– удосконалено методи прогнозування нічної гіпоглікемії у хворих на діабет першого типу і отримано кращі результати, ніж у наявних методів.

Практичне значення одержаних результатів. Запропоновано методику, за допомогою якої можна відібрати ключові значення рівня глюкози в крові з показів CGM без наявної додаткової інформації щодо схеми лікування та часу прийомів їжі, що дозволяє звести покази CGM до випадку, коли хворий міряє рівень глюкози за допомогою проб крові з пальця. Розроблено методи прогнозування нічної гіпоглікемії у хворих на діабет першого типу на основі методів машинного навчання, які дозволяють отримати кращі результати, ніж у наявних методів. Оцінено вплив демографічних даних на результати прогнозування.

Апробація результатів дисертації. Основні положення й результати роботи представлено на 18-тій міжнародній конференції SAIT 2016 (2016 р.) та VII конференції молодих вчених ПМК-2016 (2016 р.).

Публікації. Результати дисертації викладено в 3 наукових працях, у тому числі:

– VII конференція молодих вчених ПМК-2016. Тези «Прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу на основі дерев прийняття рішень»;

– VII конференція молодих вчених ПМК-2016. Тези «Методика прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу за допомогою нейронних мереж»;

– 18-та міжнародна конференція SAIT 2016. Тези «Виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу».

Ключові слова: нічна гіпоглікемія, методи машинного навчання, прогнозування, діабет 1-го типу, прогнозування глікемії на базі показань проб крові з пальця.

ABSTRACT

The thesis is presented in 73 pages. It contains 2 appendixes and bibliography of 39 references. 22 figures and 9 tables are given in the thesis.

Topic relevance. Diabetes mellitus is a chronic disease that requires constant care and supervision on the side of the patient to prevent possible complications and reduce risks of long-term complications. According to the International Diabetes Federation (IDF), worldwide there are more than 415 million people with diabetes. Hypoglycemia is a pressing problem for people with type 1 diabetes (that is, those whose body is unable to produce insulin). According to statistics, type 1 diabetes have an average of two attacks of symptomatic hypoglycemia each week and one attack of severe hypoglycemia once a year.

Nocturnal hypoglycemia prediction in patients is required to prevent drops in plasma glucose level below normal level at night. When glucose level drops below normal, functioning of the body is disrupted, which can lead to death. The creation of methods for predicting nocturnal hypoglycemia is an important task, as a result of the resolution of which could reduce the risks of life in patients with diabetes.

Thesis connection to scientific programs, plans, and topics. The thesis was prepared according to the scientific research plan of the Applied Mathematics Department of the National Technical University of Ukraine "Kyiv Polytechnic Institute."

Research goal and objectives. The goal of this thesis is to develop mathematical methods for predicting nocturnal hypoglycemia at night to prevent attacks of hypoglycemia in patients with type 1 diabetes.

To accomplish this goal, the following objectives were reached:

– develop a method of selection of the key values from glucose level time series;

– analyze existing mathematical prediction methods;

– systematize existing mathematical methods for predicting nocturnal hypoglycemia in patients with diabetes first type;

– select and adapt methods based on machine learning to solve the problem of predicting nocturnal hypoglycemia;

– create software that implements the selected machine learning methods;

– conduct experimental research of created software for clinical data of patients with diabetes first type.

Object of research is mathematical methods for prediction of nocturnal hypoglycemia for patients with type 1 diabetes

Subject of research is research and development of mathematical method of predicting nocturnal episodes of hypoglycemia in patients with type 1 diabetes mellitus based on physiological and demographic data using machine learning techniques.

Methods of research. To solve the task, the following methods were used: machine learning methods (for the development of methods for solving the problem of predicting nocturnal hypoglycemia in patients with diabetes first type); methods of the theory of algorithms and programming (for implementing the developed algorithms); methods of probability theory and mathematical statistics (for carrying out experiments).

Scientific contribution consists of the following:

– for the first time used demographic data and time measurements of blood glucose values, along with blood glucose to build forecast, unlike existing methods which use only the value of blood glucose;

– developed a method of selection the key values of blood glucose readings from time series, which were gained from CGM device;

– improved methods of predicting nocturnal hypoglycemia in patients with type 1 diabetes and obtained better results than existing techniques.

Practical value of obtained results. The method by which the key values of blood glucose from CGM readings can be selected when no additional information available regarding treatment regimens and timing of meals, which reduces CGM's time series to data available from patients who use fingerstick measurements. The methods of predicting nocturnal hypoglycemia in patients with type 1 diabetes based on machine learning

techniques, which yield better results than existing techniques. The effect of demographic data on the results of prediction.

Approbation of the thesis results. Basic ideas and results of the research were presented at the 18-th International Conference SAIT 2016 (2016) and the VII Conference of Young Scientists PMK 2016 (2016).

Publications. Thesis results are published in three scientific works:

– VII Conference of Young Scientists 2016 PMK. Abstracts "Prediction night hypoglycemia in patients with type 1 diabetes using decision trees";

– VII Conference of Young Scientists 2016 PMK. Thesis "Methods of predicting night hypoglycemia in patients with type 1 diabetes mellitus using neural networks";

– 18th International Conference SAIT 2016. Abstracts "Selection of the key values of CGM readings for predicting night hypoglycemia in patients with type 1 diabetes mellitus".

Keywords: nocturnal hypoglycemia, machine learning methods, predicting, type 1 diabetes, predicting glicemia based on fingerstick measurements.

ЗМІСТ

Перелік умовних позначень, скорочень і термінів 13

Вступ 14

1Актуальність задачі прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу 15

1.1 Хворі на діабет 1-го типу 15

1.2 Нічна гіпоглікемія 19

1.3 Постановка задачі 20

1.4 Висновки до розділу 21

2 Порівняльний аналіз існуючих методів прогнозування НГ 22

2.1 Методи прогнозування на базі показань CGM 22

2.2 Методи прогнозування на базі показань проб крові з пальця 26

2.4 Висновки до розділу 29

3 Вибір методів прогнозування нічної гіпоглікемії за результатами аналізу проб крові з пальця на цукор та з урахуванням демографічних даних 30

3.1 Критерії відбору методів прогнозування для вирішення поставленої задачі 30

3.2 Відбір груп методів прогнозування для вирішення поставленої задачі 30

3.3 Вибір методів прогнозування для вирішення поставленої задачі 38

3.4 Висновки до розділу 39

4 Виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу 40

4.1 Методика виділення ключових значень з показів CGM 40

4.2 Опис даних проекту DirecNet 46

4.3 Програмна реалізація методики 53

4.4 Результати застосування методики 56

4.5 Висновки до розділу 67

Висновки 68

Перелік посилань 69

Додаток А Лістинги програм 74

Додаток Б Ілюстративний матеріал 93

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ACC — Accuracy, точність прогнозування.

BMI — Body Mass Index, індекс маси тіла.

CART — Classification And Regression Trees, метод побудови дерев прийняття рішень.

CGM – Continious Glucose Monitoring, пристрої для неперервного зняття показів рівня глюкози у пацієнта.

F1 — F1 score, гармонічне середнє PPV і TPR.

FP (False Positive) — у матриці похибок кількість прогнозованих значень «Так», які видані, коли справжніми значеннями були «Ні» (помилка 1-го роду);

FN (False Negative) — у матриці похибок кількість прогнозованих значень «Ні», які видані, коли справжніми значеннями були «Так» (помилка 2-го роду);

HbA1C — глікірований гемоглобін, показник крові, який відображає середній вміст цукру в крові за довгостроковий період (до 3 місяців).

MCC — Matthews correlation coefficient, коефіцієнт кореляції Метью.

NPV — Negative Predictive Value, значимість негативних прогнозів.

PPV — Positive Predictive Value, значимість позитивних прогнозів.

TN (True Negative) — у матриці похибок кількість прогнозованих значень «Ні», які співпали зі справжніми значеннями «Ні».

TNR — True Negative Rate, частота негативних прогнозів.

TP (True Positive) — у матриці похибок кількість прогнозованих значень «Так», які співпали зі справжніми значеннями «Так» (вказаними у тестовій вибірці);

TPR — True Positive Rate, частота позитивних прогнозів.

Гіпоглікемія – падіння рівня глюкози в крові нижче норми (зазвичай пороговим значенням вважається 70 мг/децилітр).

Глікемія — показник рівня глюкози в крові (в мг/децилітр або ммоль/л).

НГ — нічна гіпоглікемія.

ВСТУП

Цукровий діабет 1-го типу — невиліковна хвороба, прояв якої полягає у відхиленні рівня цукру в крові від норми, що несе навантаження на організм та завдає шкоду окремим органам людини [1].

Гіпоглікемія є постійною проблемою для хворих на діабет першого типу (тобто таких, організм яких не може самостійно виробляти інсуліну). Причиною такої патології є функціональні розлади підшлункової залози, яка перестає виробляти інсулін — гормон, що регулює рівень цукру в крові [2].

Прогнозування приступів нічної гіпоглікемії у хворих є необхідним для попередження падіння рівня глюкози в плазмі крові нижче норми у нічний час доби, коли хворий не може проконтролювати свій стан. У випадку падіння рівня глюкози нижче норми, функціонування організму порушується, що може призвести до смерті. Створення методу для прогнозування нічної гіпоглікемії є актуальною задачею, в результаті вирішення якої можна зменшити ризики настання важкого стану у хворих на цукровий діабет.

У результаті дослідження проблемної області було виявлено, що наявні методи прогнозування можуть давати або короткострокові прогнози (не більше 30 хв.-2год., чого не вистачає для прогнозування нічного приступу), або потребують показів пристроїв CGM, які надто дорогі і встановлені у малої кількості хворих на діабет (у 2% від загальної кількості), або є недостатньо точними і можуть бути покращені.

Основною задачею дисертаційного дослідження є розроблення, програмна реалізація і експериментальне дослідження методів для прогнозування настання приступів нічної гіпоглікемії у хворих на цукровий діабет 1-го типу.

1 АКТУАЛЬНІСТЬ ЗАДАЧІ ПРОГНОЗУВАННЯ НІЧНОЇ ГІПОГЛІКЕМІЇ У ХВОРИХ НА ДІАБЕТ 1-ГО ТИПУ

1.1 Хворі на діабет 1-го типу

Цукровий діабет 1-го типу проявляється як відхилення рівня цукру в крові від норми, що перевантажує організм та завдає шкоду окремим органам людини [1]. Причиною такої патології є функціональні розлади підшлункової залози, яка перестає виробляти інсулін [2]. Згідно з даними International Diabetes Federation (IDF) [3], в світі нараховується більше 415 мільйонів хворих на діабет людей. Діагностується він переважно у людей дитячого та юнацького віку [1].

Організм здорової людини розщеплює отримані з їжею цукор та крохмаль до простої сполуки цукру — глюкози, що використовується організмом як енергоносій. Завдяки інсуліну клітини організму отримують глюкозу, засвоюючи її із кровотоку.

Хворі на цукровий діабет 1-го типу змушені штучно регулювати рівень цукру в крові у відповідності з чіткою схемою інсулінових ін'єкцій [4,5], дотримуючись суворої дієти [2] та проводячи планові заміри рівня глюкози в крові.

До основних факторів [1], що впливають на рівень глюкози в крові належать:

- прийом їжі хворим;
- прийом інсуліну;
- кількість та тип інсуліну, що приймає хворий;
- тип введення інсуліну: помпа чи ін'єкції;
- зріст та вага хворого;
- стать хворого;
- вік хворого;
- тривалість життя із хворобою;
- фізична активність;
- наявність синдрому Сомоджі у хворого;
- ефект «вранішньої зорі».

Основні типи інсуліну [5], що використовуються при цукровому діабеті 1-го типу, наведені разом з їхніми характеристиками в таблиці 1.1. Поширеною практикою вважається [4] комбінування в інсуліновій терапії інсуліну короткотривалої та довготривалої дії. Так, наприклад, короткотривалий інсулін призначають для компенсації наслідків прийомів їжі, а довготривалий — для симуляції нормальної роботи підшлункової залози при базальному рівні метаболізму.

Таблиця 1.1 — Основні типи інсуліну, що використовуються в інсуліновій терапії

Тип інсуліну	Застосування	Початок дії	Пік дії	Тривалість дії
NPH	Перед сном / перед вечерею	1–3 год	5–7 год	13–18 год
Lantus	Один раз в день	3–4 год	Без піку	10,8–24 год і більше
Novolog	За 5 – 10 хв перед їжею	10–15 хв	40–50 хв	3–5 год
Humalog	За 15 хв перед їжею чи одразу після їжі	5–15 хв	1–2 год	4–5 год
Regular	За 20 – 30 хв перед їжею	30–60 хв	2–4 год	6–8 год
Lente	Перед сном / перед вечерею	1–3 год	4–8 год	13–20 год
Ultralente	Перед вечерею	2–4 год	8–14 год	20–24 год

Дози та розклад прийому інсуліну визначається лікарем індивідуально для кожного пацієнта і може змінюватися в процесі лікування. Зокрема, розрахунок [6] дози короткотривалого інсуліну проводиться відповідно до мети застосування інсуліну:

$$ID_{food} = \frac{C_{total}}{R_{carb}},$$

$$ID_{correct} = \frac{GL_{current} - GL_{target}}{F_{correct}},$$

$$ID_{total} = ID_{food} + ID_{correct},$$

де ID_{food} — доза компенсації прийому їжі, $Unit$;

C_{total} — загальна кількість вуглеводів, що планується бути спожитою, $г$;

R_{carb} — кількість вуглеводів, яка компенсується одиницею інсуліну, $\frac{г}{Unit}$;

$ID_{correct}$ — доза корекції високого рівня глюкози, $Unit$;

$GL_{current}$ — поточний рівень глюкози в крові, $мг/дл$;

GL_{target} — бажаний рівень глюкози в крові, $мг/дл$;

$F_{correct}$ — зміна концентрації глюкози в крові від однієї одиниці інсуліну;

ID_{total} — сумарна доза інсуліну;

$Unit$ — величина, що характеризує кількість інсуліну; 1 $Unit$ — кількість інсуліну, необхідна для покриття фіксованої маси спожитих вуглеводів (в середньому 15 г) чи зниження рівня глюкози в крові на фіксовану різницю (в середньому 50 мг/дл).

Різні продукти по-різному впливають на рівень цукру в крові в залежності від вмісту вуглеводів та швидкості їх засвоєння [2]. У зв'язку з цим, хворі повинні чітко притримуватися дієти, в якій враховуються такі параметри їжі як глікемічний індекс (Glycemic Index, GI) та глікемічне навантаження (Glucose Load, GL).

Глікемічний індекс — показник швидкості засвоєння вуглеводів з продукту, набуває значень від 1 до 100. Природа величини — порівняння швидкості засвоєння продукту із швидкістю засвоєння такої ж маси чистої глюкози. Альтернативою глікемічному індексу виступає хлібна одиниця, зміст якої — порівняння із швидкістю засвоєння вуглеводів із такої ж маси білого хліба.

Глікемічне навантаження — показник швидкості засвоєння вуглеводів із продукту певної маси:

$$GL(A, m_A) = \frac{GI(A)}{100} * m_A * p_{CHOinA} \ ,$$

де A — тип продукту;

m_A — маса продукту A;

p_{CHOinA} — частка вуглеводів у продукті A.

Є два основні способи слідкувати за глікемією:

– аналіз проб крові з пальця на цукор;

– застосування пристроїв неперервного моніторингу глюкози (Continuous Glucose Monitor, CGM).

Аналіз проб крові з пальця на цукор — відносно дешевий спосіб слідкувати за глікемією. Щоб слідкувати за поведінкою глікемії у такий спосіб, аналіз роблять декілька разів протягом дня: перед та після основних прийомів їжі, а також перед сном. Недоліком такого підходу є часті процедури отримання крові.

Застосування пристроїв CGM для моніторингу глікемії дає повнішу картину про характер поведінки глікемії, адже пристрій отримує показання кожні 5-10 хв. Крім того, хворі не обтяжуються процедурою забору крові, оскільки пристрій вимірює глікемію неінвазивно. Але все ж, у більшості випадків, пристрій потребує періодичного калібрування, яке здійснюється за допомогою аналізу проби крові з пальця. Ще одним недоліком використання CGM є його ціна: користування пристроєм протягом двох років обходиться в середньому £2045.

1.2 Нічна гіпоглікемія

В процесі лікування, в силу певних причин, рівень глюкози в крові може відхилятись від норми й людина переходить в хворобливий стан гіперглікемії, за якого рівень цукру підіймається вище норми, чи гіпоглікемії [1], при якому рівень цукру падає нижче норми.

Гіпоглікемія супроводжується такими симптомами [7], як:

– голод;
– нервозність;
– потіння;
– запаморочення;
– сонливість;
– дезорієнтація;
– розлади мовлення;
– тривога;
– нічні кошмари.

Трапляються і важкі випадки гіпоглікемії, що призводять до коми й, навіть, смерті.

Відповідно до [8], хворі на діабет першого типу мають в середньому два приступи симптоматичної гіпоглікемії кожного тижня і один тяжкий приступ гіпоглікемії один раз на рік, що негативно позначається на рівні життя таких людей. Особливо небезпечною є гіпоглікемія, що настає вночі [9, 10], так звана нічна гіпоглікемія (НГ), в силу безпорадності хворого уві сні.

Передбачити й уникнути цього стану хворим можуть допомогти, з одної сторони, лікарі, які на основі останніх замірів рівня глюкози та особливостей хворого приймають рішення про необхідність корекції рівня цукру в крові. Але це зобов'язує хворого постійно бути під лікарським наглядом, що є незручним.

З іншої сторони, на допомогу хворим можуть прийти пристрої CGM, що в режимі реального часу знімають показники рівня цукру в крові, завдяки чому хворий може без лікаря слідкувати за своїм станом. Проте ціна таких пристроїв не по кишені більшості хворих.

1.3 Постановка задачі

Основною задачею дисертаційного дослідження є розроблення методів для прогнозування настання приступів нічної гіпоглікемії у хворих на цукровий діабет 1-го типу та їх програмна реалізація.

Розроблювані методи мають задовольняти такі вимоги:

– прогноз повинен виконуватись на невеликій кількості (не більше 8) замірів глюкози протягом дня;

– точність методу прогнозування повинна бути не нижча 75%;

– для виконання прогнозу можна також використовувати такі дані хворого, як вік, стать, зріст, вага, схема лікування, тривалість захворювання;

– результатом роботи методу має бути вердикт стосовно того, чи трапиться вночі приступ гіпоглікемії.

Додатковою задачею є отримання оцінки впливу демографічних даних про пацієнта на якість роботи розроблених методів прогнозування.

1.4 Висновки до розділу

Цукровий діабет 1-го типу — невиліковна хвороба, на яку хворіє 415 мільйонів людей, кожен з яких змушений свідомо регулювати рівень цукру у своїй крові, слідкуючи за дієтою, піддаючись інсуліновим ін'єкціям та процедурам моніторингу глікемії.

Таких хворих супроводжують часті приступи гіпоглікемії, які погіршують його самопочуття та можуть мати летальні наслідки. Особливо ж небезпечною є гіпоглікемія, що має місце вночі.

Для покращення рівня життя хворих виникає необхідність завчасного попередження гіпоглікемії. Із появою пристроїв CGM, що безперервно автоматично вимірюють та збирають дані про рівень цукру в крові, стає можливим забезпечити завчасну реакцію на відхилення рівня цукру нижче норми, проте пристрій CGM занадто дорогий, щоб його використання стало поширеною практикою.

У зв'язку з цим було поставлено задачу розробити математичні методи прогнозування нічної гіпоглікемії на основі декількох замірів рівня глюкози в крові хворого, його демографічних даних та особливостей його лікування.

2 ПОРІВНЯЛЬНИЙ АНАЛІЗ ІСНУЮЧИХ МЕТОДІВ ПРОГНОЗУВАННЯ НГ

Наявні методи прогнозування гіпоглікемії можна поділити в залежності від вхідних даних, які використовуються для побудови прогнозу, на два наступні класи :

а) методи, засновані на використанні показів CGM;

б) методи, засновані на використанні показань проб крові з пальця.

Показання CGM представляють собою часовий ряд значень глікемії, між якими часовий інтервал складає 5-15 хвилин. Це близько 96-288 замірів глюкози за день.

Проби крові з пальця не дають такої детальної інформації, оскільки це досить болісна процедура і люди в середньому роблять 7 замірів на день [11]

Варто зазначити, що на відміну від CGM, пристрої для заміру рівня глюкози в крові з пальця набагато дешевші і наявні у більшої кількості хворих на діабет, ніж CGM. Пристрої CGM встановлено лише у 2% хворих [12].

2.1 Методи прогнозування на базі показань CGM

Наразі існують наступні методи для вирішення проблеми прогнозування приступів гіпоглікемії на базі показань CGM [13-20]:

1) модифікована лінійна екстраполяція;
2) фільтр Калмана;
3) адаптивний гібридний рекурсивний фільтр;
4) статистичне прогнозування;
5) чисельний логічний алгоритм;
6) нейронні мережі;
7) метод опорних векторів;
8) моделі ARIMA (AutoRegressive Integrated Moving Average);
9) багатофакторна регресія;

10) системи нечіткого виведення.

Методи з (1) по (5) включно застосовуються для прогнозування рівня глюкози в плазмі крові в залежності від попередніх значень рівня глюкози на приблизно 30 хвилин наперед. Опис кожного окремого методу і використання їх комбінації з метою прогнозування приступів нічної гіпоглікемії наведено в [13].

Метод (1) — модифікована лінійна регресія — описується наступним чином:

$$Gl = \beta \cdot Gl_{cur} + \varepsilon,$$

де Gl — значення глюкози через 15 хв;

Gl_{cur} — поточне значення глюкози;

β — вага впливу змінної Gl_{cur};

ε — постійне зміщення.

Входом моделі слугують значення рівня глюкози. Як міра довіри прогнозу використовується значення середньоквадратичного відхилення рівня глюкози за останні 15 хвилин.

Метод (2) засновується на фільтрі Калмана. Він складається з двох кроків, які почергово повторюються [14]:

а) прогнозування значення;

б) корекція параметрів.

Фільтр використовується, щоб отримати приблизну оцінку значення глюкози і швидкості її зміни, що в подальшому дозволяє зробити прогноз стосовно рівня глюкози. У дослідженні його було налаштовано так, щоб він вносив зміни в параметри лише тоді, коли виміряна зміна глюкози відповідає реальній зміні, а не коли вона спричинена шумом у сигналі.

Метод (3) — це адаптивний гібридний рекурсивний імпульсний фільтр. Він описується наступною формулою [15]:

$$Y(n) = \frac{B(q^{-1}, n)}{A(q^{-1}, n)} \cdot x(n),$$

де $Y(n)$ — значення змінної-результата в момент часу n,

$x(n)$ — значення вхідної змінної (рівня глюкози) в момент часу n,

$A(q^{-1}, n), B(q^{-1}, n)$ — залежні від часу поліноми для оператору затримки q^{-1}.

Параметри нескінченного імпульсного фільтра постійно оновлюються відповідно до значень сигналу, зчитаних з CGM.

В основі методу (4) лежить використання статистичних моделей для прогнозування майбутнього рівня глюкози, меж помилок і ймовірності гіпоглікемії. Результат прогнозу приймається з певною мірою довіри, яка визначається експертами.

Ідея методу (5) полягає у використанні логічного виразу для винесення вердикту стосовно приступу. Спочатку розраховується зміна сигналу по трьом точкам і зчитується поточне значення глікемії. Після цього, отримані значення підставляються у логічний вираз, який прогнозує, чи відбудеться гіпоглікемія, чи ні. Цей алгоритм добре застосовувати, коли в сигналах сенсору багато викидів, адже він дозволяє їх ігнорувати.

Перевагою використання комбінації методів (1)-(5) є відносно висока точність попередження випадків гіпоглікемії (84% випадків під час дослідження). Недоліком є низький часовий поріг (не більше 35 хвилин).

Метод (6), описаний в [16], базується на використанні нейронних мереж. Як зазначають автори, метод дає точні короткострокові прогнози в денний час і точні довгострокові прогнози в нічний час (припускається, що високої точності вночі вдається досягти внаслідок низької активності хворих). Перевагою методу є точність (максимальна похибка прогнозованого значення глюкози в плазмі крові вдень при короткостроковому прогнозі на 15-60 хв складає 4,86 мг/децилітр, а при довгостроковому на 8 годин вночі — 3,6 мг/децилітр).

Метод (7) у [17] застосовували для прогнозування рівня глюкози в плазмі крові в залежності від попередніх значень рівня глюкози на 30-60 хвилин вперед. Перевагою є висока точність прогнозування (кількість попереджених випадків

гіпоглікемії більше 90%). Недоліками є складність вибору ядра, низька швидкість навчання та тестування опорних векторів.

Застосування методу (8) для задачі прогнозування значення глікемії описано в [18]. Модель ARIMA (AutoRegressive Integrated Moving Average, або інтегрована модель авторегресії — ковзного середнього) є прогностичною моделлю, яка використовує дані одновимірного часового ряду для прогнозування майбутніх значень [19]. Вона описується як:

$$\Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{j=1}^{q} b_j \varepsilon_{t-j} + \varepsilon_t,$$

де ε_t — стаціонарний часовий ряд,

c, a_i, b_j, d, p, q — параметри моделі,

Δ^d — оператор різниці часового ряду порядку d,

X_t — нестаціонарний часовий ряд.

На вхід моделі подаються рівні глюкози за останні чотири дні (навчальна вибірка) до поточного моменту часу. Результатом є 12 прогнозованих рівней глюкози з інтервалом в п’ять хвилин (тобто, прогноз стану пацієнта протягом наступної години), ґрунтуючись на яких робиться висновок чи буде приступ гіпоглікемії, чи ні. Перевагою методу є низька похибка 1-го роду ($FPR = 0{,}3\%$). Недоліками моделі є низька точність прогнозування випадків, коли гіпоглікемія матиме місце ($TPR = 9{,}9\%$). Середньоквадратичне відхилення значення глюкози при побудові прогнозу на 30 хв вперед становить 24,9 мг/децилітр, на годину вперед — 39,6 мг/децилітр.

Методи (9) і (10) застосовувалися у [20].

Багатофакторна регресія — це один з методів статистичного прогнозування, який використовує декілька спостережуваних змінних для того, щоб спрогнозувати значення змінної-результата [21]. Багатофакторна регресія описується наступною формулою:

$$Y = \sum_{i=1}^{n} \beta_i \cdot X_i + \varepsilon,$$

де Y — значення змінної-результата,

X_i — i-а спостережувана змінна,

β_i — вага впливу спостережуваної змінної X_i,

ε — постійне зміщення.

Перевагою методу (9), застосованого до задачі прогнозування гіпоглікемії, є швидкість підбору оптимальних значень вагів β_i (1 хвилина). Недоліком є низька точність побудованих прогнозів ($TPR = 51{,}78\%$, $TNR = 51{,}64\%$), у порівнянні з методом (10), застосованим до тих же даних.

Метод (10) засновується на системах нечіткого виведення. Системи нечіткого виведення використовують набори правил та функції приналежності для винесення вердикту стосовно прогнозованого значення [22]. У дослідженні [20] використовувалася система нечіткого виведення з п'ятьма функціями приналежності для кожної вхідної змінної. Перевагою є достатньо висока точність прогнозу ($TPR = 75{,}00\%$, $TNR = 51{,}64\%$). Недоліком є велика тривалість часу навчання моделі (275 хвилин).

2.2 Методи прогнозування на базі показань проб крові з пальця

Існують наступні методи для прогнозування нічної гіпоглікемії на базі показань проб крові з пальця [12, 23-25]:

1) предиктор Вінчупа-Мілнера;
2) предиктор Девіса;
3) каузальна ймовірнісна мережа;
4) LBGI (low blood glucose index);

5) лінійна комбінація існуючих предикторів.

Метод (1) ґрунтується на статистичних даних, отриманих Вінчупом і Мілнером у 1987 році [23]. Вони з'ясували, що найбільш важливим параметром для визначення — чи буде гіпоглікемія вночі, є рівень глюкози в крові перед сном. Найкращий результат в прогнозуванні дало порогове значення в 126 мг/децилітр. Даний метод основується на порівнянні значення рівня глюкози в крові пацієнта перед сном з 126 мг/децилітр. Якщо значення нижче 126 мг/децилітр, то вважається, що приступ буде, інакше — ні.

Метод (2) ґрунтується на тих же ідеях, що і метод (1), але пороговим значенням вважається 90 мг/децилітр, бо як показало дослідження Девіса, воно дозволяє попередити випадок нічної гіпоглікемії з більшим успіхом [24].

Метод (3) був запропонований у [25]. Окрім врахування рівня глюкози в крові, модель враховує дозу інсуліну і кількість вуглеводів при прийомі їжі, щоб спрогнозувати значення рівня глюкози між тестами крові. Крім того, метод (3) може бути застосовано не лише для прогнозування нічної, а й денної гіпоглікемії. Але даний метод має не надто високу надійність, адже на тестовій вибірці, на якій він прогнозував приступи гіпоглікемії протягом чотирьох ночей для 5 з 6 пацієнтів, прогноз підтвердився лише для однієї ночі з чотирьох.

Метод (4) використовує індекс LBGI для визначення чи відбудеться приступ, чи ні [12]. Значення LBGI акумулює всі денні виміри рівня глюкози і тому включає більше інформації ніж предиктор Вінкапа-Мілнера. Він визначається як:

$$LBGI = \frac{1}{n}\sum_{i=1}^{n} rl(x_i),$$

$$f(BG) = [ln(BG)^{\alpha} - \beta], \alpha, \beta > 0,$$

де $rl(BG) = 10 \cdot f(BG)^2$, якщо $f(BG) < 0$, або 0 — в інакшому випадку,

BG — рівень глюкози в крові.

α, β — параметри індексу, які залежить від границь замірів рівня глюкози в крові.

Однак, з визначення індексу LBGI виходить, що приступи можуть бути спрогнозовані лише тоді, коли пацієнт мав декілька помірно-низьких замірів глікемії, декілька дуже низьких замірів, чи суміш з обох.

Метод (5), описаний в [12], використовує лінійну комбінацію декількох предикторів з метою попередження щодо можливого нічного приступу. Як вхідні дані виступають агреговані значення в 4 точках протягом одного дня. Перевагою метода є його можливість прогнозування приступів на основі малого обсягу даних і порівняно висока точність попередження приступів ($TPR = 69{,}2\%$, $TNR = 85{,}3\%$, $PPV = 65{,}4\%$, $NPV = 87{,}4\%$, $f1 = 67{,}2\%$, $f2 = 68{,}4\%$ у наведеному дослідженні).

2.3 Наявні рішення для прогнозування гіпоглікемії

Серед рішень для виявлення гіпоглікемії на ринку існує лише один пристрій — HypoMon [26]. HypoMon — це неінвазійний пристрій, який дозволяє відслідковувати гіпоглікемію у інсулін-залежних хворих на діабет 1-го та 2-го типу. Коли пристрій прогнозує, що рівень глюкози в крові менше 45 мг/децилітр, хворий або особа, яка його доглядає, сповіщається про це. Це може використовуватися для попередження приступів нічної гіпоглікемії. В HypoMon використовується підхід на основі Баєсовських нейронних мереж, який дозволив попередити хворих в 89,2% випадків з тестової вибірки. Часовий горизонт спрацьовування пристрою не зазначається.

2.4 Висновки до розділу

Існує багато методів, які здатні спрогнозувати приступ гіпоглікемії за 30 хв. – 2 год. орієнтуючись на покази CGM з великою точністю (більше 90% випадків). Але, враховуючи поставлену у підрозділі 1.3 задачу, методи прогнозування гіпоглікемії на базі показань CGM не підходять для її вирішення. Наявні методи прогнозування на базі показань проб крові з пальця є недосконалими і можуть бути покращені.

Дане дослідження ставить метою покращення наявних результатів прогнозування НГ на базі показань проб крові з пальця.

3 ВИБІР МЕТОДІВ ПРОГНОЗУВАННЯ НІЧНОЇ ГІПОГЛІКЕМІЇ ЗА РЕЗУЛЬТАТАМИ АНАЛІЗУ ПРОБ КРОВІ З ПАЛЬЦЯ НА ЦУКОР ТА З УРАХУВАННЯМ ДЕМОГРАФІЧНИХ ДАНИХ

3.1 Критерії відбору методів прогнозування для вирішення поставленої задачі

Для відбору методів, що найкраще підходять для вирішення поставленої задачі прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу, слід враховувати такі критерії:

а) простота параметризації — має існувати простий і зрозумілий механізм регулювання параметрів методу;

б) швидкість навчання — налаштування методу на ефективну роботу не повинно займати багато часу;

в) стійкість до даних, що містять шум, — метод має бути здатен ігнорувати шум в даних;

г) здатність узагальнювати — метод має видавати правильні результати на даних, які не брали участі в процесі налаштування методу;

д) довготривалий прогноз — метод повинен видавати прогноз стосовно приступу гіпоглікемії на всю ніч, тобто горизонт прогнозування має бути близько 8 годин;

е) простота реалізації — метод має бути просто реалізувати у вигляді програми, що автоматизовано видавала би хворим прогноз стосовно приступу нічної гіпоглікемії.

3.2 Відбір груп методів прогнозування для вирішення поставленої задачі

Існують такі групи методів прогнозування:

а) методи аналізу часових рядів;

б) каузальні методи прогнозування;

в) методи експертного оцінювання;

г) методи штучного інтелекту.

Методи аналізу часових рядів [27] в процесі побудови прогнозу щодо значення певної величини у майбутньому використовують її минулі значення. До даної групи методів належать такі поширені методи, як:

а) модель середнього ковзного;

б) модель зваженого середнього ковзного (MA);

в) фільтр Калмана;

г) експоненційне згладжування;

д) модель авторегресії-ковзного середнього (ARMA);

е) інтегрована модель авторегресії-ковзного середнього (ARIMA);

ж) лінійне прогнозування.

Модель середнього ковзного та модель зваженого середнього ковзного являються одним із видів згортки і, як правило [28], використовується для згладжування короткострокових коливань та виділення основних тенденцій та циклів. Принцип побудови прогнозу за цими методами представлено формулою:

$$WWMA_t = \sum_{i=0}^{n-1} \omega_{t-i} \cdot p_{t-i} ,$$

де $WWMA_t$ — значення зваженого ковзного середнього в точці t;

n — кількість значень вихідної функції для розрахунку ковзного середнього;

ω_{t-i} — нормований ваговий коефіцієнт $t-i$-го значення вихідної функції;

p_{t-i} — значення вихідної функції в момент часу, віддалений від поточного на i інтервалів. У випадку незваженого середнього ковзного $\omega_{t-i} = 1$, $i = \overline{0, n-1}$.

В силу принципу роботи методів ковзного середнього, їх використання обмежене випадками, для яких зафіксована і наявна рівномірно розподілена в часі

статистика минулих значень величини, що прогнозується, аж до моменту в який виконується прогноз.

Фільтр Калмана — алгоритм [29], який використовує статистично зашумлені й неточні заміри з часового ряду для отримання оцінки невідомих змінних, що має вищу точність, ніж у випадку оцінки за окремими замірами.

Експоненційне згладжування [30] – метод, що прогнозує майбутнє значення невідомої змінної шляхом послідовного застосування віконної функції до всього часового ряду, що передує цьому майбутньому значенню:

$$s_t = \begin{cases} x_0, t = 0, \\ \alpha x_t + (1-\alpha)s_{t-1}, t > 0, \end{cases}$$

де s_t — значення експоненційно згладженої величини у момент часу t;

x_t — значення вихідної функції у момент часу t;

α — коефіцієнт згладжування, $0 < \alpha < 1$.

В силу своєї специфіки експоненційне згладжування використовується для прогнозування відносно стабільної чи зростаючої величини.

Модель авторегресії-ковзкого середнього (ARMA) — узагальнення [27] двох більш простих моделей часових рядів: моделі авторегресії та моделі ковзного середнього:

$$X_t = c + \epsilon_t + \sum_{i=1}^{p} \alpha_i X_{t-i} + \sum_{i=1}^{q} \beta_i \epsilon_{t-1},$$

де c — константа, $c \in \mathbb{R}$;

$\{\epsilon_t\}$ — послідовність незалежних й однаково розподілених випадкових величин з нульовим середнім (білий шум);

$\alpha_1, \dots, \alpha_p$ — авторегресійні коефіцієнти, $\alpha_1, \dots, \alpha_p \in \mathbb{R}$;

$\beta_1, \dots, \beta_q$ — коефіцієнти ковзного середнього, $\beta_1, \dots, \beta_q \in \mathbb{R}$.

Даний метод добре працює для невеликих горизонтів прогнозування величини, значення якої залежить від її попередніх значень, при чому допускається наявність шуму у вхідному часовому ряді.

Інтегрована модель авторегресії-ковзного середнього (ARIMA) — узагальнення [27] моделі ARMA з метою обробки виявлених нестаціонарних компонентів часового ряду [31]:

$$\Delta^d X_t = c + \sum_{i=1}^{p} a_i \Delta^d X_{t-i} + \sum_{j=1}^{q} b_j \epsilon_{t-j} + \epsilon_t,$$

де ϵ_t — стаціонарний часовий ряд;

c, a_i, b_j — параметри моделі;

Δ^d — оператор різниці часового ряду порядку d.

Метод лінійного прогнозування [32] прогнозує невідому величину представляючи її лінійною комбінацією попередніх значень цієї величини:

$$\hat{x}_n = \sum_{i=1}^{p} a_i x_{n-i},$$

де $\hat{x}_n$ — прогнозоване значення сигналу;

x_{n-i} — попередні значення часового ряду;

a_i — вагові коефіцієнти.

Принцип роботи даного методу, як і попередніх, нагадує принцип роботи згортки: для отримання прогнозу застосовується віконна функція до попередніх значень часового ряду. Методи такого типу можуть бути стійкими до шумів, щоправда процес підбору параметрів може виявитися складним. Крім того, дана група методів не може давати довгострокових прогнозів і узагальнювати.

Каузальні методи прогнозування орієнтовані [33] на виділення факторів, які впливають на величину, що прогнозується. До каузальних методів належить регресійний аналіз.

Регресійний аналіз — статистичний метод [34] дослідження впливу однієї чи кількох незалежних змінних $X_1, X_2, \dots, X_N$ на незалежну змінну Y. Для отримання прогнозу застосовується формула:

$$Y = b_0 + b_1 X_1 + b_2 X_2 + .. + b_N X_N,$$

де b_i — часткові коефіцієнти кореляції: $(b_i)^2$ має сенс частини дисперсії Y, що пояснюється фактором X_i.

Серед обмежень даного методу:

– погано працює при наявності кореляції між факторами і величиною, що прогнозується;

– заздалегідь має бути встановлено, що фактори впливають на величину, що прогнозується, а не навпаки;

– погано працює при відсутності кореляцій між факторами.

Дана група методів може бути використана для довготривалих прогнозів, але стійкість до шуму у них низька. Крім того, в загальному випадку, регресія може бути нелінійною і потребувати підбору функцій, що описують залежності між факторами і величиною, що прогнозується.

Методи експертного оцінювання прогнозують значення величини за допомогою поєднання думок експертів, їх суджень і суб'єктивних оцінок правдоподібності цих суджень. Такі методи зазвичай застосовують у випадках [31], коли:

– відсутні статистичних даних величини, що прогнозується;

– дані наявні, застосовуються статистичні методи прогнозування, але результат уточнюють методом експертних оцінок;

– дані наявні, незалежно виконується прогноз статистичними методами та методами експертних оцінок, їх результати комбінуються.

До методів експертного оцінювання належать:

а) метод Дельфі;

б) метод аналогій;

в) метод сценаріїв.

Зміст метода Дельфі [31] полягає в отриманні оцінок незалежних експертів та їх подальшому статистичному опрацюванні з метою виділення максимально правильного рішення.

Ідея методу аналогій полягає в припущенні, що у двох проявів певного явища закладена спільна модель поведінки [31]. Дослідження показують, що даний метод може покращувати точність прогнозу.

Метод сценаріїв [31] полягає у генерації множини можливих сценаріїв, за якими можуть розвиватися події та змінюватись величини, що прогнозуються. Ці сценарії групуються за критерієм бажаності реалізації та генерується план дій на випадок реалізації кожного сценарію.

Особливості методів експертного оцінювання [31]:

– експертна оцінка може бути неконсистентна: в силу суб'єктивності думки та природи людського мислення, що лежить в основі методу, результат роботи методу може відрізнятись в однакових експериментах;

– експертне рішення може бути спотворене корисливими мотивами;

– прогнозуючи певну величину, експерт може надавати занадто велику вагу попередньому значенню цієї величини і видавати консервативну оцінку, що призводить до системної помилки в його прогнозах.

Таким чином, методи експертного оцінювання пристосовані до довготривалих прогнозів, проте ключовим в даному методі є наявність експерта, а це не задовольняє раніш сформульований критерій простоти реалізації. Також організація прогнозу, налаштування методу — трудомісткий процес.

Методи на основі штучного інтелекту для прогнозування використовують знання про зв'язки, властивості та шаблони даних, отримані в процесі навчання.

Серед методів штучного інтелекту є такі методи прогнозування як:

а) штучні нейронні мережі;

б) метод опорних векторів;

в) методи на основі дерев прийняття рішень.

Штучні нейронні мережі вирішують сьогодні багато задач, однією з яких є прогнозування [35]. Нейронні мережі, навчаючись на прикладах, фіксують наявні тонкі функціональні зв'язки між даними, навіть якщо ці зв'язки невідомі чи важкі для опису. Особливості нейронних мереж:

– нейронні мережі здатні узагальнювати отримані знання;

– нейронні мережі являються універсальним апроксиматором;

– нейронні мережі є нелінійними, що дозволяє вирішувати задачі прогнозування часових рядів нелінійних процесів;

– нейроні мережі добре працюють із даними, що містять шум.

Метод опорних векторів [36] полягає у переведенні вхідних векторів у простір вищої розмірності і пошуку в цьому просторі гіперплощини поділу з максимальним зазором між класами. Особливості методу опорних векторів:

– стійкі до проблеми перенавчання;

– погано працюють з даними, що містять шум;

– потребують підбору функції-ядра.

Ідея методів прогнозування на основі дерев прийняття рішень [37] полягає у генерації мінімальної ієрархії правил приналежності вхідного вектору до певного класу, але достатньої для того, щоб безпомилково класифікувати інші вектори з тієї ж проблемної області. Особливості таких методів:

– не потребують підготовки даних, попередньої нормалізації;

– стійкі до викидів у даних;

– потребують регулювання глибини з метою недопущення перенавчання.

Методи на основі штучного інтелекту можуть бути просто реалізовані, вони, в основному, легко параметризуються та стійкі до шумів. Крім того, дана група методів може швидко навчатись і давати довготривалі прогнози.

Результати огляду груп методів зведено в таблиці 3.1. Отже, за встановленими критеріями для вирішення поставленої задачі прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу найкраще підходить група методів штучного інтелекту.

Таблиця 0.1 — Порівняльний аналіз груп методів прогнозування

Група методів / Критерії	Методи аналізу часових рядів	Каузальні методи прогнозування	Методи експертного оцінювання	Методи штучного інтелекту
Простота параметризації	-	-	+	+
Швидкість навчання	+	+	-	+
Стійкість до шуму	+	-	+	+
Здатність узагальнювати	-	+	+	+
Довготривалий прогноз	-	+	+	+
Простота реалізації	+	+	-	+

3.3 Вибір методів прогнозування для вирішення поставленої задачі

Порівнюючи методи штучного інтелекту (таблиця 3.2), можна виділити метод на основі штучних нейронних мереж та метод на основі дерев прийняття рішення як такі, що найбільш підходять для вирішення поставленої задачі.

Таблиця 0.2 — Порівняльний аналіз методів прогнозування з групи методів штучного інтелекту

Критерії \ Метод	Метод опорних векторів (SVM)	Метод штучних нейронних мереж (ANN)	Метод дерев прийняття рішень
Простота параметризації	-	+	+
Швидкість навчання	+	+	+
Стійкість до шуму	-	+	+
Здатність узагальнювати	+	+	+
Довготривалий прогноз	+	+	+
Простота реалізації	-	+	+

3.4 Висновки до розділу

У розділі було сформульовано критерії та проведено порівняльний аналіз методів прогнозування за цими критеріями з метою виділення найбільш підходящих для вирішення поставленої задачі прогнозування приступів нічної гіпоглікемії у хворих на діабет 1-го типу. Вибір методів було проведено в два етапи — відбір групи методів та вибір методів з відібраної групи.

У результаті першого етапу було відібрано групу методів прогнозування на основі штучного інтелекту як найбільш пристосовану до вирішення поставленої задачі.

У результаті другого етапу для подальшого дослідження було обрано метод прогнозування на основі нейронних мереж та метод прогнозування на основі дерев прийняття рішень як найбільш перспективні для вирішення поставленої задачі.

4 ВИДІЛЕННЯ КЛЮЧОВИХ ЗНАЧЕНЬ З ПОКАЗІВ CGM ДЛЯ ПРОГНОЗУВАННЯ НІЧНОЇ ГІПОГЛІКЕМІЇ У ХВОРИХ НА ЦУКРОВИЙ ДІАБЕТ 1-ГО ТИПУ

В силу поставленої в підрозділі 1.3 задачі, для прогнозування НГ мають застосовуватися показання проб крові з пальця, а не дані CGM. Оскільки наявні клінічні дослідження містять переважно часові ряди показань CGM, було вирішено розробити методику, яка дозволить звести часовий ряд до декількох ключових замірів.

Методику виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу описано в даному розділі. Її було реалізовано у вигляді програмного забезпечення засобами мов програмування R та Python і застосовано до даних проекту DirecNet.

4.1 Методика виділення ключових значень з показів CGM

Як було зазначено в розділі 2, методи прогнозування гіпоглікемії базуються або на показаннях приладів CGM, або на показаннях проб крові з пальця. Опублікована в [38] методика дозволяє виділити ключові значення рівня глюкози в крові серед часового ряду показань CGM без додаткової інформації про пацієнтів, такої як час прийому їжі чи сна. Ці ключові значення можуть бути використані для прогнозування гіпоглікемії застосовуючи методи, що базуються на показаннях проб крові з пальця.

Методика складається із таких послідовних етапів:

а) виділення ключових значень глікемії до та після прийому їжі:

1) очищення даних;

2) розбиття часового ряду на окремі дні;

3) виділення усіх значень глікемії, які відповідають замірам перед прийомом їжі та максимального впливу їжі;

4) відбір ключових значень, що відповідають першому, найбільш значущому та останньому прийомам їжі (сніданок, обід та вечеря);

5) розрахунок швидкостей зміни глікемії на основі відібраних значень.

б) виділення мінімального значення глікемії вночі.

Очищення даних проводиться з метою відкидання із часового ряду показань, що з'явилися у зв'язку з некоректною роботою пристрою CGM. Мають бути відсіяні наступні показання:

а) показання пристрою, які нижче рівня 20 мг/децилітр (надто низькі, щоб бути реалістичними);

б) різкі скачки рівня глюкози (більше 20 мг/децилітр за 5 хв).

Розбиття часового ряду на окремі дні проводиться з метою виокремлення окремих часових інтервалів, в межах яких на етапі (4) можна визначити сніданок, обід і вечерю. Інтервалами дня вважається час від 06:00 ранку дня 1 до 06:00 ранку дня 2. Такий вибір інтервалу дня дозволяє ігнорувати скачок рівня глюкози о 05:00-06:00, пов'язаний з синдромом вранішньої зорі. Окрім того, на цьому етапі відбувається відсіювання всіх днів, дані про які неповні (відсутні заміри більше ніж 1 годину протягом дня або пристрій видає постійну величину рівня глюкози протягом більш ніж 1 години).

Виділення усіх значень глікемії, які відповідають замірам перед прийомом їжі та максимального впливу їжі, необхідне для виділення ключових значень. На даному етапі розраховуються всі локальні мінімуми та максимуми протягом дня. Пари найближчих мінімумів і максимумів відповідають прийому їжі чи коливанню в результаті стресу чи шуму. Значимими вважаються такі коливання, різниця між мінімумом і максимумом яких більше 20 мг/децилітр і інтервал по часу більше 1 години (вважаємо, що вони спричинені прийомом їжі). Всі інші коливання відсіюються. Результат роботи даного етапу приведено на рисунку 4.1. Червоним точкам відповідають значення мінімумів і максимумів, які залишаються для обробки на наступному етапі роботи методики.

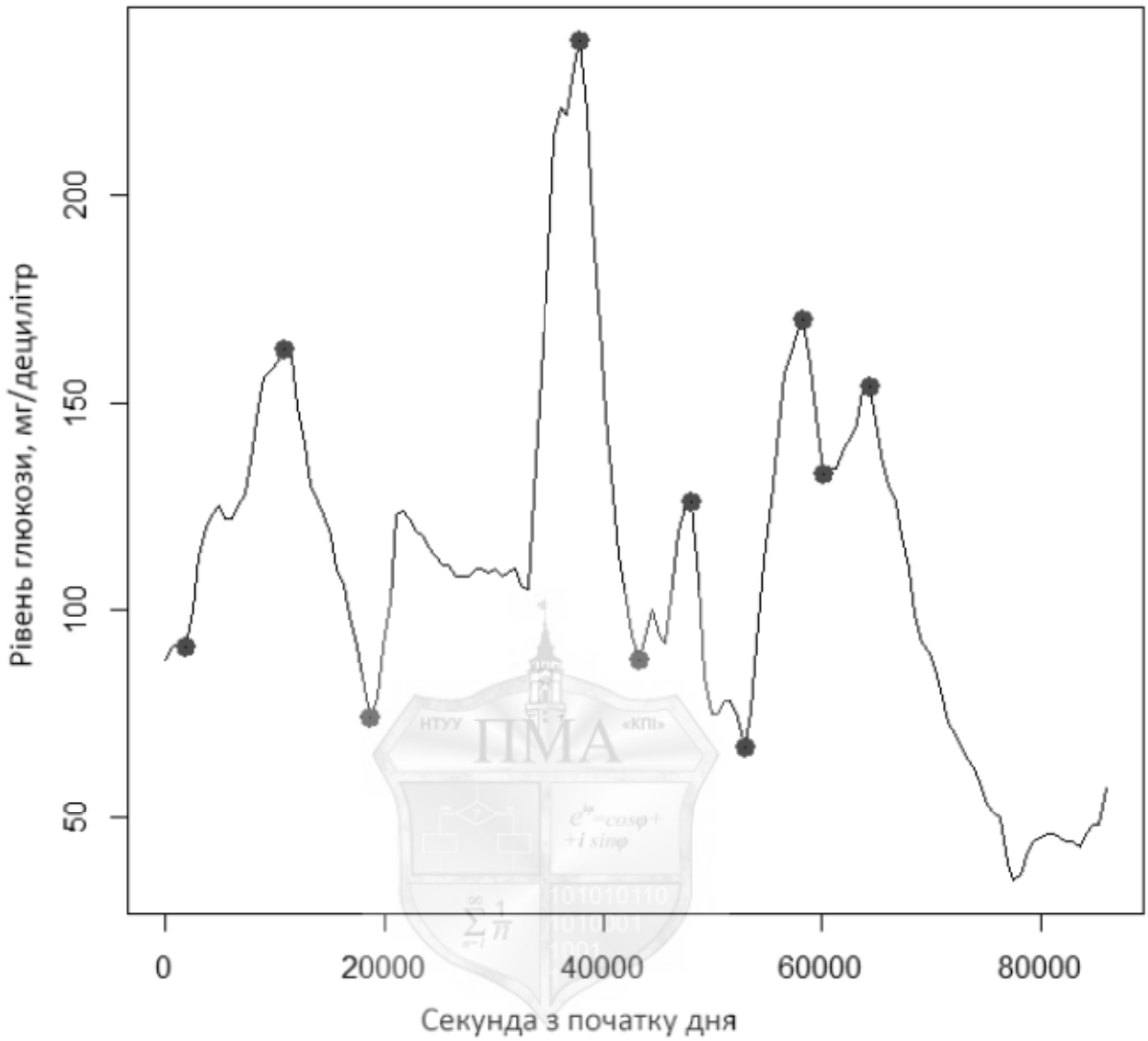

Рисунок 4.1 — Результат, отриманий після виділення усіх значень глікемії, які відповідають замірам перед прийомом їжі та максимального впливу їжі

Наступним етапом є відбір ключових значень, що відповідають першому, найбільш значущому та останньому прийомам їжі. Для цього з отриманих пар мінімальних і максимальних значень відбираються найперша пара за день (перший прийом їжі), найостанніша пара за день (останній прийом їжі) і з пар між ними така, де різниця рівня глюкози найбільша (найбільш значущий прийом їжі). Умовно ці три прийоми їжі можна позначити як сніданок, вечеря і обід відповідно. Результат роботи

даного етапу приведено на рисунку 4.2. Червоним точкам відповідають значення мінімумів і максимумів, які є ключовими.

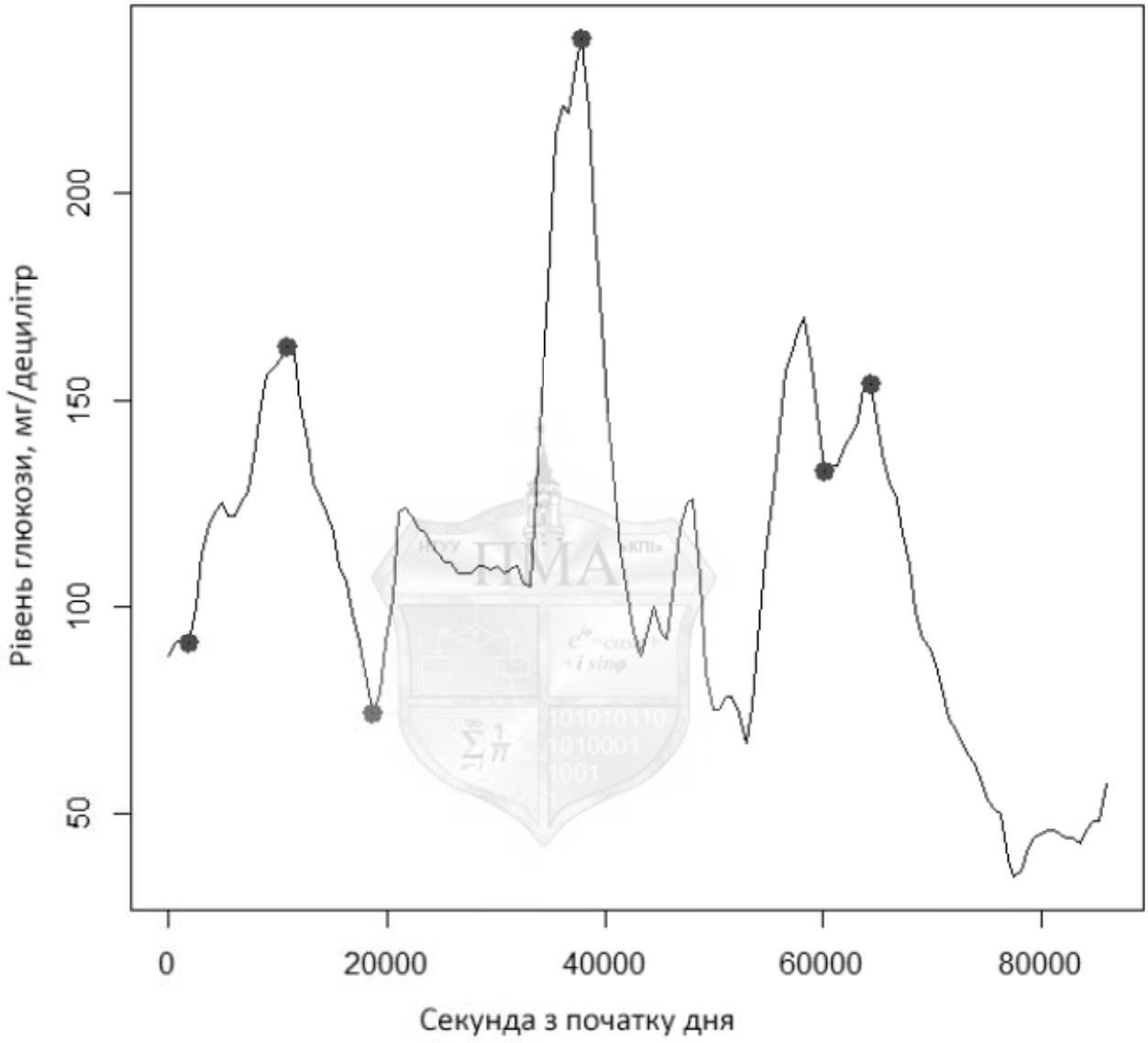

Рисунок 4.2 — Результат, отриманий після відбору ключових значень, що відповідають першому, найбільш значущому та останньому прийомам їжі

Розрахунок швидкостей зміни глікемії на основі відібраних значень відбувається використовуючи формулу кінцевої різниці між парами мінімальних і максимальних значень. Формула кінцевої різниці наступна:

$$V = \frac{Gl_{max} - Gl_{min}}{t_{max} - t_{min}},$$

де V — швидкість зміни рівня глюкози,

Gl_{min} — рівень глюкози перед прийомом їжі,

Gl_{max} — максимум рівня глюкози після прийому їжі,

t_{min} — час, коли заміряно мінімум,

t_{max} — час, коли заміряно максимум.

Результат розрахунку швидкостей росту глікемії для наведеного вище прикладу представлено в таблиці 4.1.

Таблиця 4.1 — Розрахунок швидкостей зміни глікемії на основі відібраних значень

№ п/п	Gl_{min}, мг/децилітр	Gl_{max}, мг/децилітр	t_{min}, секунд	t_{max}, секунд	V, мг/децилітр /секунду
1	91	163	1821	10834	0,007988
2	74	237	18642	37857	0,008483
3	133	154	60078	64279	0,004999

Кінцевим етапом роботи методики є визначення нічного мінімуму рівня глюкози в крові. Ніччю вважається період від останнього прийому їжі поточного дня до першого прийому їжі наступного дня. Якщо мінімальне значення складає менше 70 мг/децилітр, то треба зафіксувати, що приступ гіпоглікемії мав місце.

Результат роботи методики для тестового дня наведено в таблиці 4.2 і на рисунку 4.3. Умовні позначення:

Gl_{noct} — мінімальне значення глікемії за ніч,

t_{noct} — час, коли було зафіксовано мінімальне значення глікемії за ніч,

Gl^i_{max}, Gl^i_{min} — i-тий мінімальний і максимальний заміри глюкози відповідно.

Таблиця 4.2 — Кінцевий результат роботи методики

№ п/п	Gl_{min}, мг/децилітр	Gl_{max}, мг/децилітр	V, мг/децилітр /секунду	Gl_{noct}, мг/децилітр	t_{noct}, секунд
1	91	163	0,007988	35 (гіпоглікемія)	77499
2	74	237	0,008483		
3	133	154	0,004999		

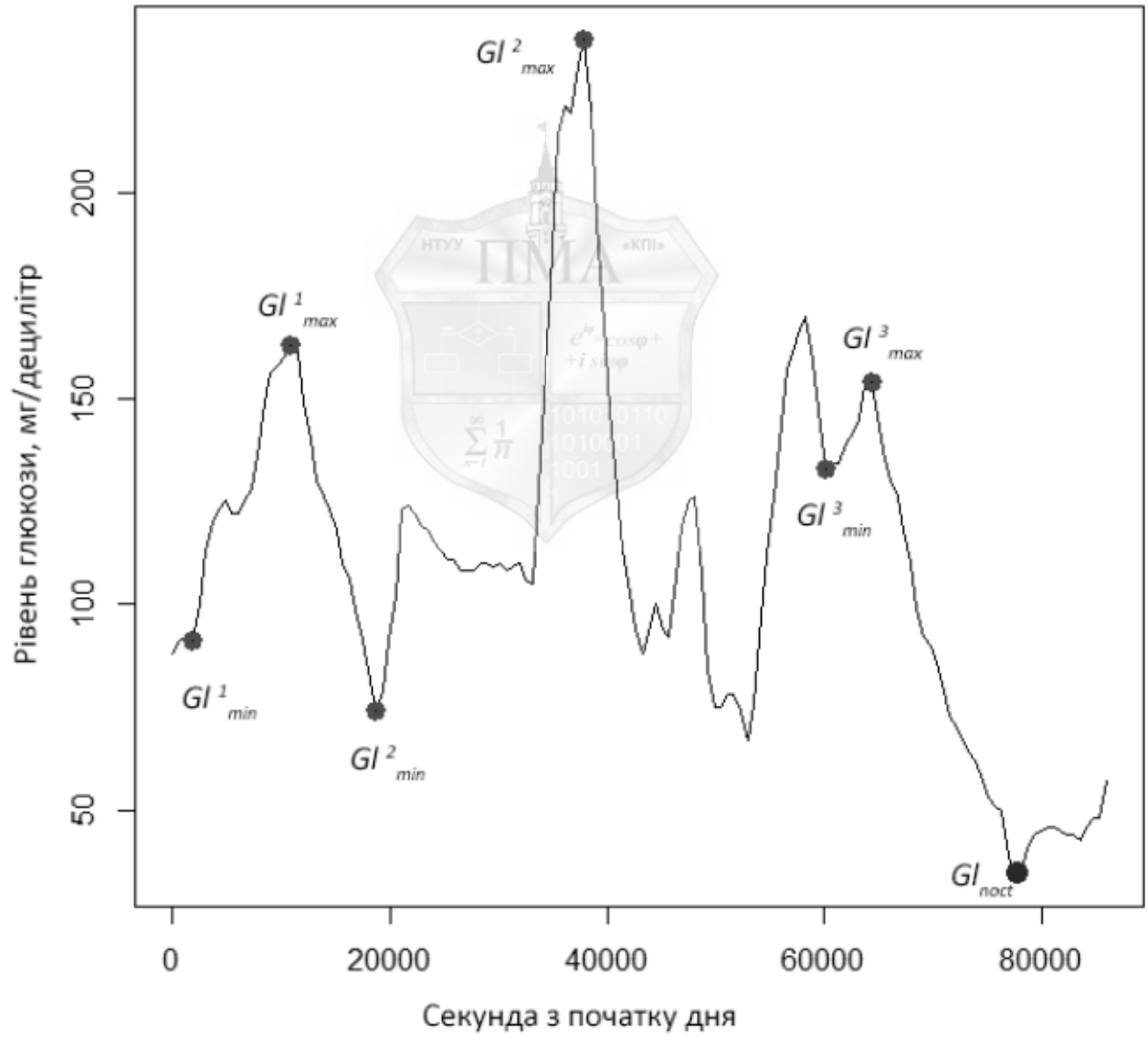

Рисунок 4.3 — Кінцевий результат роботи методики. Червоним позначено ключові значення глікемії, синім — нічний мінімум глікемії

4.2 Опис даних проекту DirecNet

Як дані для застосування методики використовувалися дані дослідження «A Pilot Study to Evaluate the Navigator Continuous Glucose Sensor in the Management of Type 1 Diabetes in Children» проекту DirecNet [39]. Ці дані було обрано через декілька причин:

а) дослідження проводилося лише для пацієнтів, хворих на діабет 1-го типу;

б) містить детальні демографічні дані стосовно кожного пацієнта;

в) велика кількість знімків рівня глюкози в крові з пристроїв CGM.

У даних дослідження міститься 798127 записів показань CGM, заміри відповідають 48 пацієнтам. Для кожного з пацієнтів наявні такі демографічні дані як:

– вік;
– стать;
– зріст;
– вага;
– дата початку захворювання;
– тривалість захворювання;
– схема лікування (ін'єкційна чи помпа, якщо ін'єкційна, то які препарати і в якому обсязі);
– рівень гемоглобіну HbA1C у хворого на початок дослідження.

Записи показань CGM для всіх пацієнтів знаходяться у файлі tblFNavGlucose.csv. Демографічні дані пацієнтів знаходяться у файлі tblFEnrollment.csv. Формати файлів наведено у таблицях 4.3 та 4.4 відповідно.

Таблиця 4.3 — Формат файлу з показаннями CGM (tblFNavGlucose.csv)

Назва поля	Опис
RecID	Номер запису
PtID	Ідентифікатор пацієнта
NavReadDt	Дата зчитування рівня глюкози
NavReadTm	Час зчитування рівня глюкози
Gl	Рівень глюкози

Таблиця 4.4 — Формат файлу з демографічними даними пацієнтів (tblFEnrollment.csv)

Назва поля	Опис	Мінімум	Максимум	Можливі значення
RecID	Номер запису			
PtID	Ідентифікатор пацієнта			
VisitDt	Дата запису на програму			
EligVer	Підходить для дослідження			1 => Так
Gender	Стать			M => Чоловік, F => Жінка
Ethnicity	Етнос			Hispanic or Latino, Not Hispanic or Latino, Unknown/not reported

Продовження таблиці 4.4.

Назва поля	Опис	Мінімум	Максимум	Можливі значення
Race	Раса			White, Black/African American, Asian, Native Hawaiian/Other Pacific Islander, American Indian/Alaskan Native, More than one race, Unknown/not reported
RaceDs	Суміш рас			
OnsetDt	Дата прийняття на облік в лікарні			
NPH	Застосовуваний інсулін: NPH			1 => Застосовує
Lente	Застосовуваний інсулін: Lente			1 => Застосовує
UltraLente	Застосовуваний інсулін: UltraLente			1 => Застосовує
Lantus	Застосовуваний інсулін: Lantus			1 => Застосовує
Novolog	Застосовуваний інсулін: Novolog			1 => Застосовує
Humalog	Застосовуваний інсулін: Humalog			1 => Застосовує

Продовження таблиці 4.4.

Назва поля	Опис	Мінімум	Максимум	Можливі значення
Regular	Застосовуваний інсулін: Regular			1 => Застосовує
InsOth	Застосовуваний інсулін: Інший			1 => Застосовує
InsOthDs	Застосовуваний інсулін: Опис іншого інсуліну			
DShot	Кількість ін'єкцій в день	0		
FLenPumpUse	Тривалість використання інсулінової помпи			6 mon -<1 yr, 1-<2 yrs, 2-<5 yrs, >=5 yrs
PumpType	Тип інсулінової помпи			Smart Pump, Regular Pump
InsCarbB	Відношення інсуліну до кількості вуглеводів під час сніданку: одиниць на грам вуглеводів			
InsCarbBNotUsed	Поле InsCarbB не використовується			1 => Так
InsCarbL	Відношення інсуліну до кількості вуглеводів під час ланчу: одиниць на грам вуглеводів			
InsCarbLNotUsed	Поле InsCarbL не використовується			1 => Так

Продовження таблиці 4.4.

Назва поля	Опис	Мінімум	Максимум	Можливі значення
InsCarbD	Відношення інсуліну до кількості вуглеводів під час обіду: одиниць на грам вуглеводів			
InsCarbDNotUsed	Поле InsCarbD не використовується			1 => Так
InsCarbBS	Відношення інсуліну до кількості вуглеводів перед сном: одиниць на грам вуглеводів			
InsCarbBSNotUsed	Поле InsCarbBS не використовується			1 => Так
UsualInsDoseB	Звичайна додаткова доза інсуліну під час сніданку	0		
UsualInsDoseL	Звичайна додаткова доза інсуліну під час ланчу	0		
UsualInsDoseD	Звичайна додаткова доза інсуліну під час обіду	0		
UsualInsDoseS	Звичайна додаткова доза інсуліну під час перекусу	0		

Продовження таблиці 4.4.

Назва поля	Опис	Мінімум	Максимум	Можливі значення
UsualInsDoseBS	Звичайна додаткова доза інсуліну перед сном	0		
AvgCorFactmgdl	Середній показник корекції чутливості	10	500	
AvgCorFactAbmgdl	Середній показник корекції чутливості	100	500	
AvgCorFactNotUsed	Середній показник корекції чутливості не використовується			1 => Так
NumSevHypo	Кількість випадків втрачання свідомості за останні 6 місяців внаслідок гіпоглікемії			0, 1, 2, 3, >3
PriorCGMUse	Prior continuous glucose monitor use			Так, Ні
CGMS	Раніше використовував CGMS			1 => Так
GWB	Раніше використовував GWB			1 => Так
CGMOther	Раніше використовував інший спосіб моніторингу глікемії			1 => Так
CGMOtherDs	Опис іншого способу моніторингу глікемії			

Продовження таблиці 4.4.

Назва поля	Опис	Мінімум	Максимум	Можливі значення
EduCareGvr1	Опікун: Мама, тато, інший			Mother, Father, Other
EduCareGvr1a	Рівень освіти опікуна			<4, 4, 5, 6, 7, 8, 9, 10, 11, 12, Associates, Bachelors, Masters, Professional
OthCareGvr	Інший опікун			Grandmother, Grandfather, Aunt, Uncle, Older Sibling
EduCareGvr2	Опікун: Мама, тато			Mother, Father
EduCareGvr2a	Рівень освіти опікуна			<4, 4, 5, 6, 7, 8, 9, 10, 11, 12, Associates, Bachelors, Masters, Professional
PEExamDt	День огляду			
Weight	Вага, кг	10	180	
Height	Зріст, см	30	210	
HbA1CDt	Дата тестування			
HbA1C	Показник HbA1C (DCA2000)	4	15	

4.3 Програмна реалізація методики

Методику, описану в підрозділі 4.1, було реалізовано у вигляді програмного забезпечення засобами мов програмування R і Python. Схему роботи програмного забезпечення наведено на рисунку 4.4. Модулі, з яких складається програмна реалізація методики, і їх опис наведено у таблиці 4.5. Вихідні тексти модулів наведено у додатку А.

Таблиця 4.5 — Модулі програмної реалізації методики виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу

Модуль	Опис
PatientSorter.py	Скрипт, який розділює таблицю tblFNavGlucose.xlsx даних дослідження [direcnet web link] по окремим пацієнтам і неперервним проміжкам записів CGM для цих пацієнтів, і зберігає окремі .xlsx файли з іменем, який має формат виду Patient#<Id>\Period_from_<DateStart>.xlsx, де Id — ідентифікатор пацієнта, DateStart — дата початку неперервного проміжку.
charts-generator.py	Скрипт, який будує графіки в файлах формату Patient#<Id>\Period_from_<DateStart>.xlsx, для того, щоб можна було легко переглянути коливання рівня глюкози в крові в програмі Microsoft Excel.
build-plot.r	Функції для побудови графіків по окремим дням з файлів виду Patient#<Id>\Period_from_<DateStart>.xlsx і збереження їх у форматі PNG.

Продовження таблиці 4.5.

Модуль	Опис
config.r	Налаштування шляхів і підключення необхідних бібліотек для коректної роботи модулів, написаних на R.
count-stats.r	Підрахунок статистики стосовно кількості днів замірів по кожному окремому пацієнту і сумарно.
create-csv.r	Функції, які застосовують методику виділення ключових значень CGM до окремих пацієнтів чи всіх одразу.
dataset-processing.r	Функції для виділення окремого дня з неперервного проміжку записів у файлах виду Patient#<Id>\Period_from_<DateStart>.xlsx, розрахунку відліку в секундах від початку дня, переліку всіх наявних папок з файлами пацієнтів та визначення нічного мінімуму.
filter-data-v2.r	Функції для виокремлення екстремумів серед часового ряду записів CGM та виділення ключових записів серед екстремумів.
main.r	Головний файл програми, який викликає функції з інших модулей з метою застосування методики виділення ключових значень CGM до даних, отриманих після обробки скриптом PatientSorter.py.

Рисунок 4.4 — Схема роботи програми, що реалізує методику виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу

4.4 Результати застосування методики

У результаті застосування методики було відібрано сумарно 1088 доби замірів 48 пацієнтів, які використовували пристрої CGM. До результуючого файлу було додано демографічні дані. Повний опис формату результуючого файлу наведено у таблиці 4.6.

Таблиця 4.6 — Опис таблиці, отриманої в результаті застосування методики і додавання демографічних даних

Назва поля	Опис
PtId	Ідентифікатор пацієнта
DT_Fix	Дата дня
BMax_T1	Час заміру перед прийомом сніданку
BMax_Gl1	Значення глікемії перед прийомом сніданку
Max_T1	Час заміру на піку впливу їжі після сніданку
Max_Gl1	Значення глікемії на піку впливу їжі після сніданку
BMax_T2	Час заміру перед прийомом обіду
BMax_Gl2	Значення глікемії перед прийомом обіду
Max_T2	Час заміру на піку впливу їжі після обіду
Max_Gl2	Значення глікемії на піку впливу їжі після обіду
BMax_T3	Час заміру перед прийомом вечері
BMax_Gl3	Значення глікемії перед прийомом вечері
Max_T3	Час заміру на піку впливу їжі після вечері
Max_Gl3	Значення глікемії на піку впливу їжі після вечері
TheLastBeforeBed	Значення глікемії за 8 годин до першого прийому їжі наступного дня (вважаємо, що це замір перед сном)
NoctMin_T	Час мінімального значення вночі
NoctMin_Gl	Рівень мінімального значення глюкози вночі

Продовження таблиці 4.6.

Назва поля	Опис
Hypoglycemia	Чи спостерігалася гіпоглікемія в цю ніч
Ill_years	Тривалість захворювання в роках
Age	Вік
Gender	Стать
Height	Зріст, см
Weight	Вага, кг
InsMod	Схема лікування (ін'єкції чи помпа)
V1	Швидкість зміни глікемії після сніданку
V2	Швидкість зміни глікемії після обіду
V3	Швидкість зміни глікемії після вечері
HbA1C	Рівень гемоглобіну HbA1C на момент прийому в програму
BMI	Індекс маси тіла

Файл з даними в цьому форматі у подальшому використовується для тестування та навчання моделей прогнозування нічної гіпоглікемії.

На основі отриманих даних було побудовано гістограми, які кажуть про розподіл пацієнтів у вибірці (рисунки 4.5-4.12) та відносну частоту випадків гіпоглікемії в залежності від різних факторів (рисунки 4.13-4.20). Засобами математичного пакету STATISTICA побудовано матрицю кореляції числових атрибутів (рисунки 4.21 і 4.22). З рисунка 4.22 можна зробити висновок, що найбільш значимий вплив на нічний рівень глюкози має останній замір глюкози перед сном.

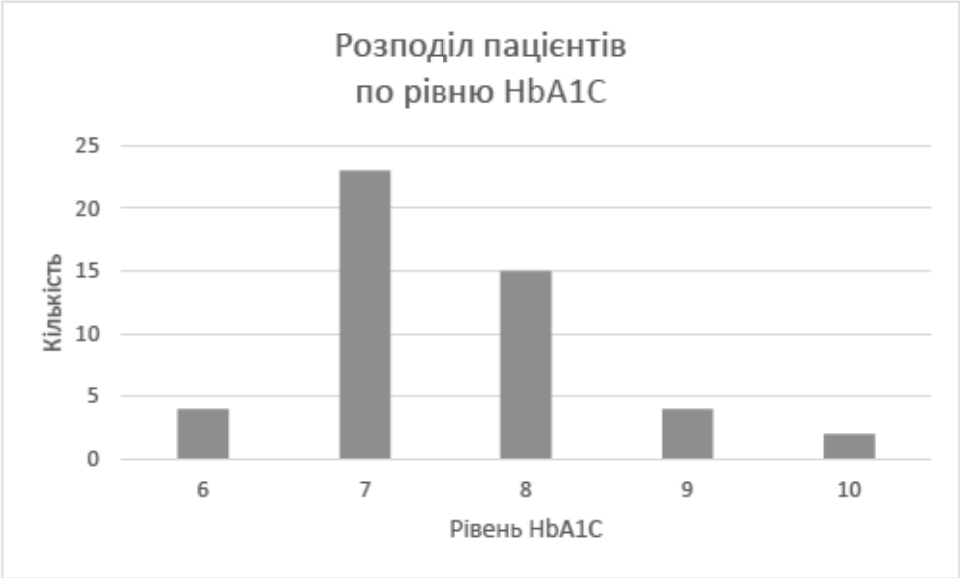

Рисунок 4.5 — Розподіл пацієнтів по рівню HbA1C

Рисунок 4.6 — Розподіл пацієнтів по статі

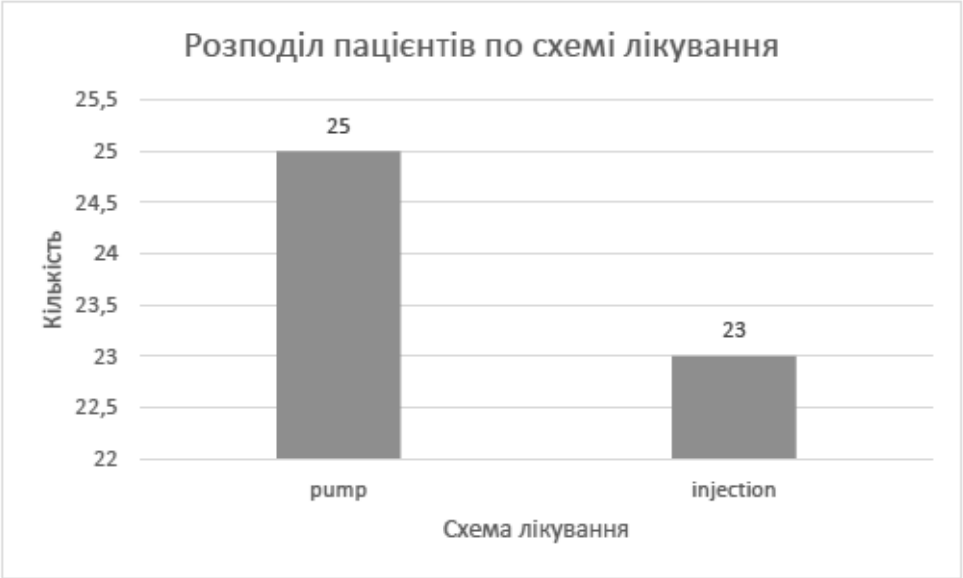

Рисунок 4.7 — Розподіл пацієнтів по схемі лікування

Рисунок 4.8 — Розподіл пацієнтів по віку

Рисунок 4.9 — Розподіл пацієнтів по зросту

Рисунок 4.10 — Розподіл пацієнтів по вазі

Рисунок 4.11 — Розподіл пацієнтів по індексу маси тіла

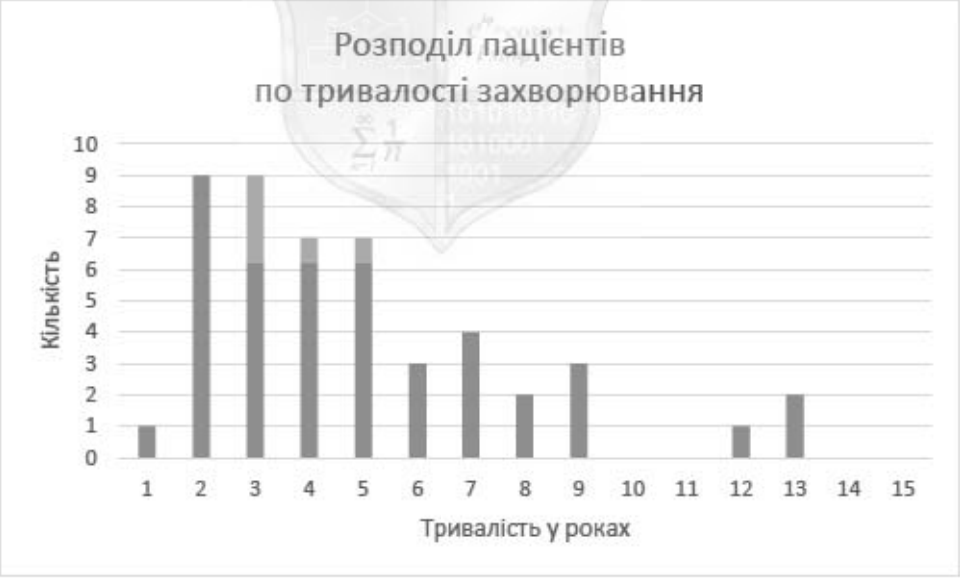

Рисунок 4.12 — Розподіл пацієнтів по тривалості захворювання

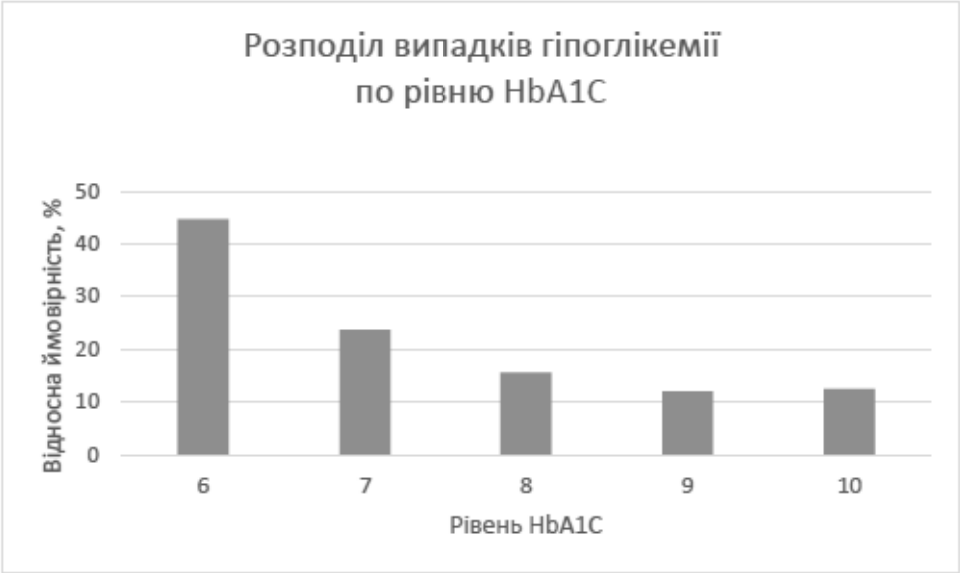

Рисунок 4.13 — Розподіл випадків гіпоглікемії по рівню HbA1C

Рисунок 4.14 — Розподіл випадків гіпоглікемії по статі

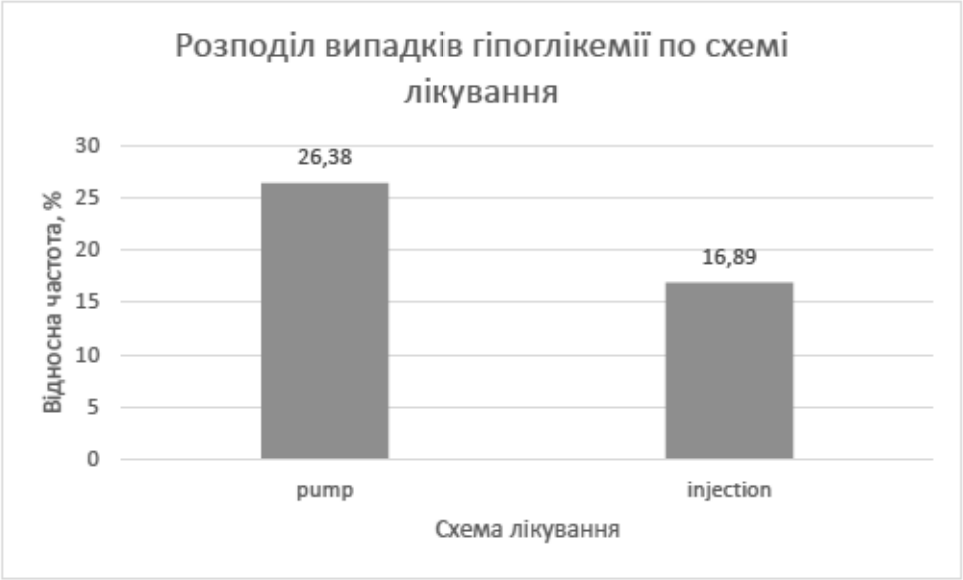

Рисунок 4.15 — Розподіл випадків гіпоглікемії по схемі лікування

Рисунок 4.16 — Розподіл випадків гіпоглікемії по віку

Рисунок 4.17 — Розподіл випадків гіпоглікемії по зросту

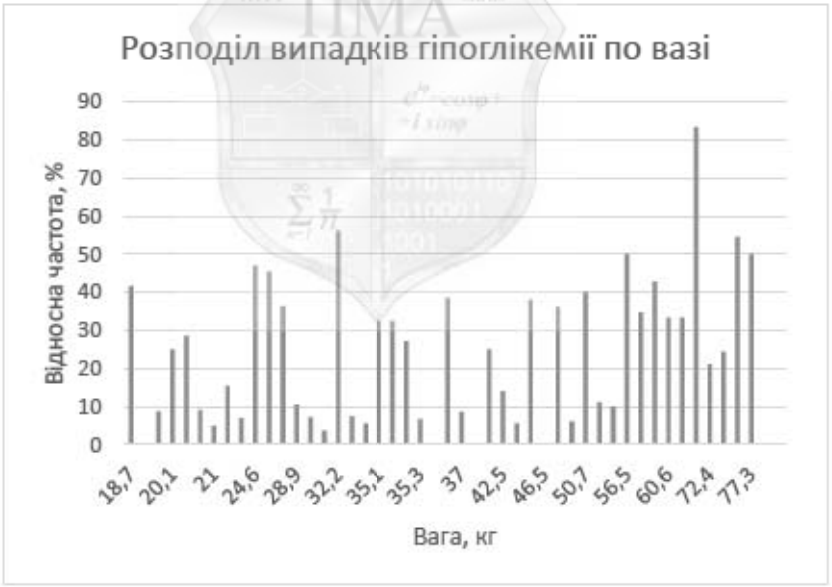

Рисунок 4.18 — Розподіл випадків гіпоглікемії по вазі

Рисунок 4.19 — Розподіл випадків гіпоглікемії по індексу маси тіла

Рисунок 4.20 — Розподіл випадків гіпоглікемії по тривалості захворювання

Correlations (all.sta)
Casewise deletion of MD
N=1088

Variable	BMax_T1	BMax_GI1	Max_T1	Max_GI1	BMax_T2	BMax_GI2	Max_T2	Max_GI2	BMax_T3	BMax_GI3	Max_T3
BMax_T1	1,00	-0,13	0,74	-0,10	0,33	-0,03	0,31	-0,06	0,11	0,05	0,03
BMax_GI1	-0,13	1,00	-0,18	0,41	-0,06	0,22	-0,04	0,16	-0,06	0,13	-0,03
Max_T1	0,74	-0,18	1,00	0,09	0,48	0,08	0,42	-0,10	0,14	0,05	0,03
Max_GI1	-0,10	0,41	0,09	1,00	0,15	0,27	0,12	0,24	-0,02	0,23	-0,09
BMax_T2	0,33	-0,06	0,48	0,15	1,00	-0,10	0,88	-0,16	0,38	0,11	0,20
BMax_GI2	-0,03	0,22	0,08	0,27	-0,10	1,00	-0,13	0,46	-0,09	0,21	-0,07
Max_T2	0,31	-0,04	0,42	0,12	0,88	-0,13	1,00	0,00	0,48	0,21	0,27
Max_GI2	-0,06	0,16	-0,10	0,24	-0,16	0,46	0,00	1,00	0,18	0,42	0,05
BMax_T3	0,11	-0,06	0,14	-0,02	0,38	-0,09	0,48	0,18	1,00	0,07	0,68
BMax_GI3	0,05	0,13	0,05	0,23	0,11	0,21	0,21	0,42	0,07	1,00	-0,16
Max_T3	0,03	-0,03	0,03	-0,09	0,20	-0,07	0,27	0,05	0,68	-0,16	1,00
Max_GI3	-0,01	0,15	-0,04	0,26	-0,10	0,26	-0,08	0,31	-0,34	0,48	-0,15
Ill_years	0,00	-0,08	0,01	-0,13	-0,09	-0,02	-0,06	-0,04	0,01	-0,11	0,01
Age	0,03	-0,06	0,12	-0,09	-0,00	0,03	-0,00	-0,11	-0,01	-0,06	-0,06
Height	0,01	-0,05	0,11	-0,11	-0,02	0,03	-0,02	-0,13	-0,01	-0,09	-0,04
Weight	-0,01	-0,06	0,10	-0,13	-0,02	0,03	-0,03	-0,13	-0,02	-0,08	-0,05
V1	0,13	-0,28	-0,12	0,24	-0,03	-0,06	-0,02	0,11	0,02	0,07	-0,02
V2	-0,00	-0,03	-0,04	0,07	0,11	-0,21	-0,12	0,17	0,05	0,01	-0,01
V3	0,05	-0,01	0,06	0,17	0,05	0,07	0,03	0,13	0,00	-0,03	-0,35
HbA1C	-0,00	0,16	-0,03	0,21	-0,05	0,10	-0,03	0,22	-0,07	0,21	-0,01
BMI	-0,01	-0,09	0,08	-0,13	-0,03	0,01	-0,04	-0,12	-0,03	-0,06	-0,07
TheLastBeforeBed	0,03	0,11	0,03	0,19	0,02	0,22	0,07	0,29	-0,00	0,41	0,15
NoctMin_T	-0,06	0,04	-0,06	-0,05	0,02	-0,02	0,04	-0,00	0,17	-0,10	0,30
NoctMin_GI	-0,03	0,21	-0,05	0,18	-0,00	0,21	0,01	0,21	0,03	0,18	0,22

Рисунок 4.21 — Матриця кореляції числових атрибутів. Частина 1 з 2

Variable	Max_GI3	Ill_years	Age	Height	Weight	V1	V2	V3	HbA1C	BMI	TheLastBeforeBed	NoctMin_T	**NoctMin_GI**
BMax_T1	-0,01	0,00	0,03	0,01	-0,01	0,13	-0,00	0,05	-0,00	-0,01	0,03	-0,06	-0,03
BMax_GI1	0,15	-0,08	-0,06	-0,05	-0,06	0,28	-0,03	-0,01	0,16	-0,09	0,11	0,04	0,21
Max_T1	-0,04	0,01	0,12	0,11	0,10	-0,12	-0,04	0,06	-0,03	0,08	0,03	-0,06	-0,05
Max_GI1	0,26	-0,13	-0,09	-0,11	-0,13	0,24	0,07	0,17	0,21	-0,13	0,19	-0,05	0,18
BMax_T2	-0,10	-0,09	-0,00	-0,02	-0,02	-0,03	0,11	0,05	-0,05	-0,03	0,02	0,02	-0,00
BMax_GI2	0,26	-0,02	0,03	0,03	0,03	-0,06	-0,21	0,07	0,10	0,01	0,22	-0,02	0,21
Max_T2	-0,08	-0,06	-0,00	-0,02	-0,03	-0,02	-0,12	0,03	-0,03	-0,04	0,07	0,04	0,01
Max_GI2	0,31	-0,04	-0,11	-0,13	-0,13	0,11	0,17	0,13	0,22	-0,12	0,29	-0,00	0,21
BMax_T3	-0,34	0,01	-0,01	-0,01	-0,02	0,02	0,05	0,00	-0,07	-0,03	-0,00	0,17	0,03
BMax_GI3	0,48	-0,11	-0,06	-0,09	-0,08	0,07	0,01	-0,03	0,21	-0,06	0,41	-0,10	0,18
Max_T3	-0,15	0,01	-0,06	-0,04	-0,05	-0,02	-0,01	-0,35	-0,01	-0,07	0,15	0,30	0,22
Max_GI3	1,00	-0,16	-0,13	-0,18	-0,18	0,11	0,04	0,32	0,28	-0,14	0,74	-0,06	0,38
Ill_years	-0,16	1,00	0,27	0,26	0,25	-0,08	-0,08	-0,08	-0,04	0,17	-0,10	0,03	-0,09
Age	-0,13	0,27	1,00	0,95	0,89	-0,07	-0,07	-0,06	-0,18	0,63	-0,13	0,02	-0,09
Height	-0,18	0,26	0,95	1,00	0,94	-0,09	-0,08	-0,09	-0,20	0,63	-0,16	0,01	-0,09
Weight	-0,18	0,25	0,89	0,94	1,00	-0,10	-0,09	-0,07	-0,17	0,84	-0,17	0,02	-0,13
V1	0,11	-0,08	-0,07	-0,09	-0,10	1,00	0,10	0,13	0,06	-0,09	0,08	-0,05	0,02
V2	0,04	-0,08	-0,07	-0,08	-0,09	0,10	1,00	0,12	0,10	-0,09	-0,00	-0,03	0,03
V3	0,32	-0,08	-0,05	-0,09	-0,07	0,13	0,12	1,00	0,03	-0,05	0,19	-0,09	-0,01
HbA1C	0,28	-0,04	-0,18	-0,20	-0,17	0,06	0,10	0,03	1,00	-0,07	0,25	0,02	0,29
BMI	-0,14	0,17	0,63	0,63	0,84	-0,09	-0,09	-0,05	-0,07	1,00	-0,16	0,03	-0,16
TheLastBeforeBed	0,74	-0,10	-0,13	-0,16	-0,17	0,08	-0,00	0,19	0,25	-0,16	1,00	-0,04	0,44
NoctMin_T	-0,06	0,03	0,02	0,01	0,02	-0,05	-0,03	-0,09	0,02	0,03	-0,04	1,00	0,11
NoctMin_GI	0,38	-0,09	-0,09	-0,09	-0,13	0,02	0,03	-0,01	0,29	-0,16	0,44	0,11	1,00

Рисунок 4.22 — Матриця кореляції числових атрибутів. Частина 2 з 2

4.5 Висновки до розділу

В рамках дослідження було розроблено методику виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу і затосовано до даних дослідження «A Pilot Study to Evaluate the Navigator Continuous Glucose Sensor in the Management of Type 1 Diabetes in Children» проекту DirecNet.

Розроблена методика дозволяє отримати ключові значення рівнів глюкози в крові з часового ряду показів CGM, коли невідомо точний час прийому їжі чи сна пацієнта, тобто звести дані CGM до випадку, коли хворий міряє рівень глюкози за допомогою взяття проб крові з пальця. Важливою перевагою методики є те, що вона може бути автоматизована.

Програмна реалізація методики була створена засобами мови програмування R та Python. На вхід програми подається файл з показами рівня глюкози в крові, які знімав пристрій CGM та файл з демографічними даними. В результаті роботи програми створюється новий файл з даними, які відібрані відповідно до описаної методики.

З матриці кореляції числових атрибутів (рисунки 4.21 та 4.22) виявлено, що найбільш значимий вплив на нічний рівень глюкози має, насамперед, останній замір глюкози перед сном.

Отримані за допомогою методики дані у подальшому використовуються для прогнозування нічної гіпоглікемії методами штучного інтелекту.

ВИСНОВКИ

У дисертаційній роботі одержано такі нові теоретичні та практичні результати:

а) на основі аналізу проблемної області було поставлено задачу розроблення математичних методів прогнозування нічної гіпоглікемії на основі декількох замірів рівня глюкози в крові хворого, його демографічних даних та особливостей його лікування і визначено обмеження на ці методи;

б) проведено аналіз існуючих методів прогнозування і визначено формальні критерії розв'язання поставленої в підрозділі 1.3 задачі. Відповідно до визначених критеріїв, в якості методів прогнозування обрано методи штучного інтелекту (машинного навчання);

в) запропоновано методику, за допомогою якої можна відібрати ключові значення рівня глюкози в крові з показів CGM без наявної додаткової інформації щодо схеми лікування та часу прийомів їжі, що дозволяє звести покази CGM до випадку, коли хворий міряє рівень глюкози за допомогою проб крові з пальця;

г) програмно реалізовано методику виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу;

д) розроблену методику застосовано до даних проекту DirecNet, які у подальшому використовуються під час розробки методів прогнозування нічної гіпоглікемії для порівняння результатів роботи методів;

е) оцінено розподіл пацієнтів у вибірці даних проекту DirecNet, відповідно до значень кожного з атрибутів;

ж) проведено кореляційний аналіз демографічних і фізіологічних факторів. Виявлено, що найбільше на показник того, чи відбудеться приступ нічної гіпоглікемії, впливає значення глюкози перед сном.

У подальшому методику виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу може бути розширено за рахунок автоматизації виділення значення останнього заміру рівня глюкози перед сном.

ПЕРЕЛІК ПОСИЛАНЬ

1. American Diabetes Association [Електронний ресурс] : [Веб-сайт] / [American Diabetes Association]. — Електронні дані. — [Alexandria], 1995-2016. — Режим доступу: http://www.diabetes.org/ (дата звернення 08.06.2016). — Назва з екрану.

2. Harvard T.H. Chan School of Public Health [Електронний ресурс] : [Веб-сайт]. — Електронні дані. — [Boston], 2016. — Режим доступу: https://www.hsph.harvard.edu (дата звернення 08.06.2016). — Назва з екрану.

3. IDF Diabetes Atlas / International Diabetes Federation. — 7 ed. — Brussels, Belgium: International Diabetes Federation, 2015. — 144 p.

4. Hirsch Irl B. Type 1 Diabetes Mellitus and the Use of Flexible Insulin Regimens / Irl B. Hirsch // American Family Physician. — Seattle, Washington : University of Washington School of Medicine, 1999. — Vol. 60, № 8. — P. 2343-2352.

5. WebMD : Diabetes Health Center : Types of Insulin for Diabetes Treatment [Електронний ресурс] : [Стаття]. — Електронні дані. — 2015. — Режим доступу: http://www.webmd.com/diabetes/guide/diabetes-types-insulin (дата звернення 08.06.2016). — Назва з екрану.

6. Advanced Insulin Management: Using Insulin-to-Carb Ratios and Correction Factors [Електронний ресурс]: [Стаття] / eatright.org. Academy of Nutrition and Dietetics. — Електронні дані. — Режим доступу: http://www.wcu.edu/WebFiles/PDFs/6403AdvancedInsulinManagementFinal.pdf (дата звернення 08.06.2016) . — Назва з екрану.

7. NIH National Institute of Diabetes and Digestive and Kidney Diseases : Hypoglycemia [Електронний ресурс] : [Стаття] / [The National Institute of Diabetes and Digestive and Kidney Diseases]. — Електронні дані. — [Bethesda], 2008. — Режим доступу: http://www.niddk.nih.gov/health-information/health-

topics/Diabetes/hypoglycemia/Pages/index.aspx (дата звернення 08.06.2016). — Назва з екрану.

8. Shafiee G. The importance of hypoglycemia in diabetic patients / G. Shafiee, M. MohajeriTehrani, M. Pajouhi, B. Larijani // Journal of Diabetes & Metabolic Disorders. — 2012. — Vol. 11, № 8. — P. 1-7.

9. Cryer P. The barrier of hypoglycemia in diabetes // Diabetes. — 2008. — Vol. 57, № 12. — P. 3169-3176.

10. Майоров А.Ю. Клинические и психологические аспекты гипогликемии при сахарном диабете / А.Ю. Майоров, О.Г. Мельникова // Сахарный диабет. — 2010. — №3. — С. 46-50.

11. Kuenen J. HbA1c results in relation to familiar every-day measurements-the near future / J. Kuenen, R. Borg and ADAG Study Group // Diabetes Voice. — Belgica, 2009. — Vol. 54, № 1. — P. 33-36.

12. Tkachenko P. Prediction of Nocturnal Hypoglycemia by an aggregation of previously known prediction approaches: Proof of concept for clinical application / P. Tkachenko, G. Kriukova, M. Aleksandrova, O. Chertov, E. Renard, S. Pereverzyev [Електронний ресурс]. — Режим доступу: http://www.ricam.oeaw.ac.at/files/reports/16/rep16-06.pdf

13. Buckingham B. Prevention of nocturnal hypoglycemia using predictive alarm algorithms and insulin pump suspension / B. Buckingham, H.P. Chase, E. Dassau, E. Cobry, P. Clinton, V. Gage, K. Caswell, J. Wilkinson, F. Cameron, H. Lee, B.W. Bequette // Diabetes Care. — 2010. — Vol. 33 (5). — P. 1013-1017.

14. Welch G. An introduction to the kalman filter / G. Welch, G. Bishop // Proceedings of the Siggraph Course. — Los Angeles, 2001. — 81 p.

15. Williamson G. Digital Signal Processing Handbook / G. A. Williamson, Ed. Vijay K. Madisetti and Douglas B. Williams. — Boca Raton: CRC Press LLC, 1999. — 21 p.

16. Robertson G. Blood glucose prediction using artificial neural networks trained with the AIDA diabetes simulator: a proof-of-concept pilot study / G. Robertson, E.D.

Lehmann, W. Sandham and D. Hamilton // Journal of Electrical and Computer Engineering. — 2011. — P. 2-14.

17. Georga E. A glucose model based on support vector regression for the prediction of hypoglycemic events under free-living conditions / E.I. Georga, V.C. Protopappas, D. Ardigo, D. Polyzos and D.I. Fotiadis // Diabetes technology & therapeutics. — 2013. — Vol. 15 (8). — P. 634-643.

18. Plis K. A machine learning approach to predicting blood glucose levels for diabetes management / K. Plis, R. Bunescu, C. Marling, J. Shubrook & F. Schwartz // Modern Artificial Intelligence for Health Analytics. Papers from the AAAI-14. — 2014. — P. 35-39.

19. Stage, Forecasting, and Using ARIMA Procedure Statements. The ARIMA Procedure [Електронний ресурс]. — Режим доступу: http://www.okstate.edu/sas/v8/saspdf/ets/chap7.pdf

20. San P. P. A novel extreme learning machine for hypoglycemia detection / P.P. San, S.H. Ling, N.N. Soe, H.T. Nguyen // In Engineering in Medicine and Biology Society (EMBC), 36th Annual International Conference of the IEEE. — 2014. — P. 302-305.

21. Kutner M. Applied linear statistical models / Michael H. Kutner … [et al.]. — 5th ed. — Irwin: McGraw-Hill, 1996. — 1415 p.

22. Ross T. Applied linear statistical models / Timothy J. Ross. — 3rd ed. — Chichester: John Wiley & Sons, Ltd, 2010. — 585 p.

23. Whincup G. Prediction and management of nocturnal hypoglycaemia in diabetes / G. Whincup, R. Milner // Archives of Disease in Childhood. — 1987. — Vol. 62 (4). — P. 333-337.

24. Davies A. Prediction and management of nocturnal hypoglycaemia in diabetes // Archives of Disease in Childhood. — 1987. — Vol. 62 (10). — P. 1085.

25. Cavan D. Use of the {DIAS} model to predict unrecognised hypoglycaemia in patients with insulin-dependent diabetes / D. Cavan, R. Hovorka, O. Hejlesen, S. Andreassen, P. Snksen // Computer Methods and Programs in Biomedicine. — 1987. — Vol. 50 (3). — P. 241-246.

26. HypoMon for detection of hypoglycaemia in diabetics [Електронний ресурс]. — Режим доступу: http://www.horizonscanning.gov.au/internet/horizon/publishing.nsf/Content/BB580B674729F620CA2575AD0080F351/$File/HypoMon%20for%20detection%20of%20hypoglycaemia%20in%20diabetics.pdf

27. Бокс Дж. Анализ временных рядов. Прогноз и управление / Дж. Бокс, Г. Дженкинс. — М. : Мир, 1974. — Т. 1. — 406 с.; М. : Мир, 1974. — Т. 2. — 194 с.

28. Грешилов А. А. Математические методы построения прогнозов. / А. А. Грешилов, В. А. Стакун, А. А. Стакун. — М. : Радио и связь, 1997. — 112 с.

29. Kalman R. E. A New Approach to Linear Filtering and Prediction Problems / R. E. Kalman // Transactions of the ASME – Journal of Basic Engineering. — 1960. — Vol. 82, Series D. — P. 35-45.

30. Gardner E. S. Exponential smoothing: The state of the art / E. S. Gardner // Journal of Forecasting. — 1985. — Vol. 4, Issue 1. — P. 1-28.

31. Hyndman R. J. Forecasting: principles and practice / R. J. Hyndman, G. Athanasopoulos. — Otexts, 2013. — 292 p.

32. Солонина А. И. Основы цифровой обработки сигналов: курс лекций. — 2-е. / А. И. Солонина, Д. А. Улахович, С. М. Арбузов, Е. Б. Соловьева. — СПб : БХВ-Петербург, 2005. — 765 с.

33. Nahmias S. Production and Operations Analysis / S. Nahmias. — New York: McGraw-Hill/Irwin, 2009. — 789 p.

34. Armstrong J. S. Illusions in Regression Analysis / J. S. Armstrong // International Journal of Forecasting. — 2012. — Vol. 28, № 3. — P. 689.

35. Guoqiang Z. Forecasting with artificial neural networks: The state of the art / Z. Guoqiang, B. E. Patuwo, Y. Hu. Michael // International journal of forecasting. — 1998. — Vol. 14, № 1. — P. 35-62.

36. Lijuan C. Support vector machines experts for time series forecasting / C. Lijuan // Neurocomputing. — 2003. — Vol. 51. — P. 321-339.

37. Breiman L. B. Classification and regression trees / L. B. Breiman, J. H. Friedman, R. A. Olshen, C. J. Stone. — Monterey, CA: Wadsworth & Brooks/Cole Advanced Books & Software, 1984. — 368 p.

38. Чертов О. Виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу / О.Р. Чертов, С.Ю. Сахаров, Д.В. Юрченко // Системний аналіз та інформаційні технології : матеріали 18-ї Міжнародної науково-технічної конференції SAIT 2016. — К.: УНК "ИПСА" НТУУ "КПИ", 2016. — С. 176-177.

39. A Pilot Study to Evaluate the Navigator Continuous Glucose Sensor in the Management of Type 1 Diabetes in Children // Diabetes Research in Children Network (DirecNet) [Електронний ресурс]. — Режим доступу: http://direcnet.jaeb.org/Studies.aspx?RecID=166

НАЦІОНАЛЬНИЙ ТЕХНІЧНИЙ УНІВЕРСИТЕТ УКРАЇНИ

«КИЇВСЬКИЙ ПОЛІТЕХНІЧНИЙ ІНСТИТУТ»

Факультет прикладної математики

Кафедра прикладної математики

«На правах рукопису»

УДК 519.2:616.379-008.64

«До захисту допущено»

Завідувач кафедри

_____ О. Р. Чертов

«___» _____ 2016 р.

Магістерська дисертація

на здобуття ступеня магістра

зі спеціальності 8.04030101 «Прикладна математика»

на тему: Прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу. Метод із застосуванням дерев прийняття рішень

Виконав: студент II курсу, групи КМ-41м

Юрченко Дмитро Володимирович _____

Науковий керівник зав. кафедри, д-р техн. наук, доцент _____

Чертов О. Р.

Консультант із нормоконтролю старший викладач Мальчиков В. В. _____

Рецензент зав. відділу, д-р техн. наук, доцент _____

Валькман Ю. Р.

Засвідчую, що в цій магістерській дисертації немає запозичень із праць інших авторів без відповідних посилань.

Студент _____

Київ — 2016

Національний технічний університет україни

«Київський політехнічний інститут»

Факультет прикладної математики

Кафедра прикладної математики

Рівень вищої освіти — другий (магістерський)

Спеціальність 8.04030101 «Прикладна математика»

ЗАТВЕРДЖУЮ

Завідувач кафедри

_____ О. Р. Чертов

«___» _____ 2016 р.

ЗАВДАННЯ

на магістерську дисертацію студенту

Юрченку Дмитру Володимировичу

1. Тема дисертації: «Прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу. Метод із застосуванням дерев прийняття рішень»,
науковий керівник дисертації Чертов Олег Романович, д-р техн. наук, доцент,
затверджені наказом по університету від «21» березня 2016 р. № 1187-С.
2. Термін подання студентом дисертації: «10» червня 2016 р.
3. Об'єкт дослідження: математичні методи прогнозування нічної гіпоглікемія у хворих на діабет 1-го типу.
4. Предмет дослідження: розробка та дослідження математичного методу прогнозування нічних приступів гіпоглікемії у хворих на цукровий діабет 1-го типу на основі фізіологічних і демографічних показників на основі дерев прийняття рішень.
5. Перелік завдань, які потрібно розробити: проаналізувати існуючі математичні методи прогнозування на основі дерев прийняття рішень, обрати та пристосувати методи на основі дерев прийняття рішень для вирішення задачі прогнозування нічної гіпоглікемії, створити програмне забезпечення, що реалізує обрані методи, провести

експериментальне дослідження створеного програмного забезпечення на клінічних даних хворих на діабет першого типу.

6. Орієнтовний перелік ілюстративного матеріалу: приклад дерева прийняття рішень, екранні форми програмної реалізації, порівняльна таблиця результатів роботи методів, графік впливу значень атрибутів на результати прогнозування.

7. Орієнтовний перелік публікацій: тези «Виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу», тези «Прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу на основі дерев прийняття рішень».

8. Дата видачі завдання: «2» березня 2016 р.

Календарний план

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
1	Ґрунтовне ознайомлення з предметною областю	15.12.2014	
2	Визначення структури магістерської дисертації; вивчення літератури, пошук додаткової літератури	01.03.2015	
3	Робота над першим, другим та третім розділами спільної частини магістерської дисертації	15.05.2015	
4	Проведення наукового дослідження; робота над четвертим розділом спільної частини магістерської дисертації	15.10.2015	
5	Проведення наукового дослідження; робота над статтею за результатами наукового дослідження	15.12.2015	

№ з/п	Назва етапів виконання магістерської дисертації	Термін виконання етапів магістерської дисертації	Примітка
6	Робота над першим, другим, третім та четвертим розділами індивідуальної частини магістерської дисертації; підготовка статті за результатами наукового дослідження; розроблення програмного забезпечення	01.03.2016	
7	Завершення роботи над основною частиною магістерської дисертації	15.05.2016	
8	Оформлення текстової і графічної частин магістерської дисертації	25.05.2016	

Студент _____ Юрченко Д. В.

Науковий керівник дисертації _____ Чертов О. Р.

РЕФЕРАТ

Дисертацію виконано на 51 аркуш, вона містить 1 додаток та перелік посилань на використані джерела з 22 найменувань. У роботі наведено 6 рисунків та 8 таблиць.

Актуальність теми. Цукровий діабет — це хронічне захворювання, яке потребує постійного медичного догляду і нагляду з боку самого хворого, щоб попередити можливі ускладнення та зменшити ризик довгострокових ускладнень. Згідно з даними International Diabetes Federation (IDF), в світі нараховується більше 415 мільйонів хворих на діабет людей. Гіпоглікемія є нагальною проблемою для хворих на діабет першого типу (тобто таких, організм котрих не може самостійно виробляти інсуліну). Відповідно до статистики, хворі на діабет першого типу мають в середньому два приступи симптоматичної гіпоглікемії кожного тижня і один тяжкий приступ гіпоглікемії один раз на рік.

Прогнозування приступів нічної гіпоглікемії у хворих є необхідним для попередження падіння рівня глюкози в плазмі крові нижче норми у нічний час доби. У випадку падіння рівня глюкози нижче норми, функціонування організму порушується, що може призвести до смерті. Тому створення методів прогнозування нічної гіпоглікемії є важливою задачею, в результаті вирішення якої можна зменшити ризики для життя хворих на цукровий діабет.

Зв'язок роботи з науковими програмами, планами, темами. Дисертаційна робота виконувалась згідно з планом науково-дослідних робіт кафедри прикладної математики Національного технічного університету України «Київський політехнічний інститут».

Мета і задачі дослідження. Метою дисертаційної роботи є розробка математичних методів прогнозування нічної гіпоглікемії для попередження приступів нічної гіпоглікемії у хворих на діабет першого типу.

Для досягнення вказаної мети було розв'язано такі задачі:

– проаналізувати існуючі математичні методи прогнозування на основі дерев прийняття рішень;

– обрати та пристосувати методи на основі дерев прийняття рішень для вирішення задачі прогнозування нічної гіпоглікемії;

– створити програмне забезпечення, що реалізує обрані методи;

– провести експериментальне дослідження створеного програмного забезпечення на клінічних даних хворих на діабет першого типу.

Об'єктом дослідження є математичні методи прогнозування нічної гіпоглікемія у хворих на діабет 1-го типу.

Предметом дослідження є розробка та дослідження математичного методу прогнозування нічних приступів гіпоглікемії у хворих на цукровий діабет 1-го типу з урахуванням фізіологічних і демографічних показників на основі дерев прийняття рішень.

Методи дослідження. Для розв'язання поставленої задачі використовувалися такі методи: методи дерев прийняття рішень (для розроблення методів розв'язання задачі прогнозування нічної гіпоглікемії у хворих на діабет першого типу); методи теорії алгоритмів та програмування (для програмної реалізації розроблених алгоритмів); методи теорії ймовірності та математичної статистики (для аналізу результатів експериментів).

Наукова новизна одержаних результатів складається з наступних положень:

– уперше використано демографічні дані та час замірів рівня глюкози в крові разом зі значеннями рівня глюкози в крові для побудови прогнозу, на відміну від існуючих методів, де використовується лише значення рівня глюкози в крові;

– удосконалено методи прогнозування нічної гіпоглікемії у хворих на діабет першого типу і отримано кращі результати, ніж у наявних методів.

Практичне значення одержаних результатів. Розроблено методи прогнозування нічної гіпоглікемії у хворих на діабет першого типу на основі дерев прийняття рішень, які дозволяють отримати кращі результати, ніж у наявних методів. Оцінено вплив демографічних даних на результати прогнозування.

Апробація результатів дисертації. Основні положення й результати роботи представлено на 18-тій міжнародній конференції SAIT 2016 (2016 р.) та VII конференції молодих вчених ПМК-2016 (2016 р.).

Публікації. Результати дисертації викладено в 2 наукових працях, а саме:

– VII конференція молодих вчених ПМК-2016. Тези «Прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу на основі дерев прийняття рішень»;

– 18-та міжнародна конференція SAIT 2016. Тези «Виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу».

Ключові слова: нічна гіпоглікемія, методи машинного навчання, прогнозування, діабет 1-го типу, прогнозування глікемії на базі показань проб крові з пальця, дерева прийняття рішень, CART, C4.5, Random Forest, Boosting, AdaBoost.

ABSTRACT

The thesis is presented in 51 pages. It contains one appendix and bibliography of 22 references. Six figures and eight tables are given in the thesis.

Topic relevance. Diabetes mellitus is a chronic disease that requires constant care and supervision on the side of the patient to prevent possible complications and reduce risks of long-term complications. According to the UN World Health Organization (WHO), worldwide there are more than 415 million people with diabetes. Hypoglycemia is a pressing problem for people with type 1 diabetes (that is, those whose body is unable to produce insulin). According to statistics, type 1 diabetes have an average of two attacks of symptomatic hypoglycemia each week and 1 attack of severe hypoglycemia once a year.

Nocturnal hypoglycemia prediction in patients is required to prevent drops in plasma glucose level below normal level at night. When glucose level drops below normal, functioning of the body is disrupted, which can lead to death. The creation of methods for predicting nocturnal hypoglycemia is an important task, as a result of the resolution of which could reduce the risks of life in patients with diabetes.

Thesis connection to scientific programs, plans, and topics. The thesis was prepared according to the scientific research plan of the Applied Mathematics Department of the National Technical University of Ukraine "Kyiv Polytechnic Institute."

Research goal and objectives. The goal of this thesis is to develop mathematical methods for predicting nocturnal hypoglycemia at night to prevent attacks of hypoglycemia in patients with type 1 diabetes.

To accomplish this goal, the following objectives were reached:

– analyze existing mathematical prediction methods;

– select and adapt methods based on decision trees to solve the problem of predicting nocturnal hypoglycemia;

– create software that implements the selected machine learning methods;

– conduct experimental research of created software for clinical data of patients with diabetes first type.

Object of research is mathematical methods for prediction of nocturnal hypoglycemia for patients with type 1 diabetes

Subject of research is research and development of mathematical method of predicting nocturnal episodes of hypoglycemia in patients with type 1 diabetes mellitus based on physiological and demographic data using decision trees techniques.

Methods of research. To solve the task, the following methods were used: decision trees methods (for the development of methods for solving the problem of predicting nocturnal hypoglycemia in patients with diabetes first type); methods of the theory of algorithms and programming (for implementing the developed algorithms); methods of probability theory and mathematical statistics (for carrying out experiments).

Scientific contribution consists of the following:

– for the first time used demographic data and time measurements of blood glucose values, along with blood glucose to build forecast, unlike existing methods which use only the value of blood glucose;

– improved methods of predicting nocturnal hypoglycemia in patients with type 1 diabetes and obtained better results than existing techniques.

Practical value of obtained results. The methods of predicting nocturnal hypoglycemia in patients with type 1 diabetes based on decision trees techniques, which yield better results than existing techniques. The effect of demographic data on the results of prediction.

Approbation of the thesis results. Basic ideas and results of the research were presented at the 18-th International Conference SAIT 2016 (2016) and the VII Conference of Young Scientists PMK 2016 (2016).

Publications. Thesis results are published in two scientific works:

– VII Conference of Young Scientists 2016 PMK. Abstracts "Prediction night hypoglycemia in patients with type 1 diabetes using decision trees";

– 18th International Conference SAIT 2016. Abstracts "Selection of the key values of CGM readings for predicting night hypoglycemia in patients with type 1 diabetes mellitus".

Keywords: nocturnal hypoglycemia, machine learning methods, predicting, type 1 diabetes, predicting glicemia based on fingerstick measurements, decision trees, CART, C4.5, Random Forest, Boosting, AdaBoost.

ЗМІСТ

Перелік умовних позначень, скорочень і термінів 13

Вступ 14

1 Прогнозування на основі дерев прийняття рішень 15

1.1 Класифікація методів прогнозування на основі дерев прийняття рішень 16

1.2 Алгоритм ID3 20

1.3 Алгоритм C4.5 20

1.4 Алгоритм CART 21

1.5 Алгоритм CHAID 22

1.6 Метод Bagging 23

1.7 Сімейство методів Random Forest 24

1.8 Метод Boosting 25

1.9 Висновки до розділу 26

2 Реалізація методу прогнозування НГ на базі дерев прийняття рішень 28

2.1 Особливості застосування методу на підготовлених даних з проекту DirecNet 28

2.2 Програмна реалізація методу 30

2.3 Висновки до розділу 34

3 Результати експериментального дослідження 35

3.1 Метрики порівняння результатів роботи методів 35

3.2 Набори даних для тестування методів 37

3.3 Результати роботи моделей 39

3.4 Аналіз впливу факторів на результат прогнозування 45

3.5 Висновки до розділу 48

Висновки 49

Перелік посилань 50

Додаток А Лістинги програм 52

ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

ACC — Accuracy, точність прогнозування.

BMI — Body Mass Index, індекс маси тіла.

CART — Classification And Regression Trees, метод побудови дерев прийняття рішень.

CGM – Continious Glucose Monitoring, пристрої для неперервного зняття показів рівня глюкози у пацієнта.

F1 — F1 score, гармонічне середнє PPV і TPR.

FP (False Positive) — у матриці похибок кількість прогнозованих значень «Так», які видані, коли справжніми значеннями були «Ні» (помилка 1-го роду);

FN (False Negative) — у матриці похибок кількість прогнозованих значень «Ні», які видані, коли справжніми значеннями були «Так» (помилка 2-го роду);

HbA1C — глікірований гемоглобін, показник крові, який відображає середній вміст цукру в крові за довгостроковий період (до 3 місяців).

MCC — Matthews correlation coefficient, коефіцієнт кореляції Метью.

NPV — Negative Predictive Value, значимість негативних прогнозів.

PPV — Positive Predictive Value, значимість позитивних прогнозів.

TN (True Negative) — у матриці похибок кількість прогнозованих значень «Ні», які співпали зі справжніми значеннями «Ні».

TNR — True Negative Rate, частота негативних прогнозів.

TP (True Positive) — у матриці похибок кількість прогнозованих значень «Так», які співпали зі справжніми значеннями «Так» (вказаними у тестовій вибірці);

TPR — True Positive Rate, частота позитивних прогнозів.

Гіпоглікемія – падіння рівня глюкози в крові нижче норми (зазвичай пороговим значенням вважається 70 мг/децилітр).

Глікемія — показник рівня глюкози в крові (в мг/децилітр або ммоль/л).

НГ — нічна гіпоглікемія.

ВСТУП

Після застосування методики виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу на даних проекту DirecNet, що було описано у розділі 4 спільної частини, отримано набір даних, ґрунтуючись на якому здійснено розробку та експериментальне дослідження методів прогнозування нічної гіпоглікемії на основі дерев прийняття рішень.

У наступних розділах піде мова про наявні методи прогнозування на основі дерев прийняття рішень і методи покращення результатів прогнозування. Після цього описано процес пристосування методів прогнозування до поставленої задачі, їх параметри і програмну реалізацію, яка була виконана за допомогою засобів мови програмування R. Отримані моделі було протестовано, результати їх роботи порівняно і зведено до порівняльної таблиці.

Основною задачею дисертаційного дослідження є розроблення методів для прогнозування настання приступів нічної гіпоглікемії у хворих на цукровий діабет 1-го типу на основі дерев прийняття рішень та їх програмна реалізація.

Розроблювані методи мають задовольняти такі вимоги:

– прогноз повинен виконуватись на невеликій кількості (не більше 8) замірів глюкози протягом дня;

– точність методу прогнозування повинна бути не нижча 75%;

– для виконання прогнозу можна також використовувати такі дані хворого, як вік, стать, зріст, вага, схема лікування, тривалість захворювання;

– результатом роботи методу має бути вердикт стосовно того, чи трапиться вночі приступ гіпоглікемії.

Додатковою задачею є отримання оцінки впливу демографічних даних про пацієнта на якість роботи розроблених методів прогнозування.

1 ПРОГНОЗУВАННЯ НА ОСНОВІ ДЕРЕВ ПРИЙНЯТТЯ РІШЕНЬ

Задача класифікації, яка ставить метою віднесення об'єктів до певних заздалегіть визначених класів, є розповсюдженою і виникає у різноманітних сферах науки і техніки [1]. Прикладом таких задач є визначення чи є електронні повідомлення спамом в залежності від заголовку повідомлення і тексту, визначення клітин як злоякісних чи нормальних в залежності від результатів знімку МРТ, і класифікація галактик в залежності від їх форми.

Вхідними даними для будь-якої задачі класифікації є набір записів. Кожен запис, також відомий як приклад чи екземпляр, характеризується кортежем $(\bar{x}, y)$, де $\bar{x}$ — це набір атрибутів, а y — особливий атрибут, відомий як цільовий атрибут. Набір атрибутів $\bar{x}$ може включати в себе як дискретні, так і неперервні значення, а цільовий атрибут y приймає лише дискретні значення. Це ключовий момент, який відрізняє задачу класифікації від регресії, що є задачею прогнозування в якій y є неперервним значенням.

Іншими словами, класифікація — це задача визначення цільової функції $f: \bar{x} \to y$, де $\bar{x}$ — це набір атрибутів, який відображується на цільовий атрибут y.

Задача класифікації може використовуватися для описового моделювання та прогностичного моделювання. Дерева прийняття рішень можуть використовуватися для вирішення обох задач, але нас цікавить саме прогностичне моделювання значення гіпоглікемії на основі фізіологічних та демографічних даних, тобто визначення заздалегіть невідомого значення цільового атрибуту y в залежності від значень набору атрибутів $\bar{x}$.

1.1 Класифікація методів прогнозування на основі дерев прийняття рішень

Дерево прийняття рішень — це класифікатор, який визначається як рекурсивне розбиття простору екземплярів [2]. Дерево прийняття рішень складається з вузлів, які формують дерево з коренем, тобто таке, у якого є лише один «кореневий» елемент, що не має жодних вхідних дуг. Всі інші вузли мають рівно одну вхідну дугу. Вузол, з якого виходить дуга, називається внутрішнім або тестовим вузлом. Всі інши вузли називаються листовими (також відомі як термінальні вузли або вузли-рішення). В дереві прийняття рішень, кожен внутрішній вузол розділює простір екземплярів на два або більше підпростори в залежності від певної дискретної функції значень вхідних атрибутів. В найпростішому і найбільш розповсюдженому випадку, кожен внутрішній вузол враховує лише один атрибут, так, що простір екземплярів розбивається відповідно до його значення. У випадку чисельних значень, умові відповідає проміжок значень.

Кожному листовому значенню відповідає лише один клас, який співвідносить набір вхідних даних до значення цільового атрибуту. Можливий також варіант, коли лист включає в себе ймовірнісний вектор, який містить ймовірності того, що цільовий атрибут прийме певне значення. Екземпляри класифікуються, розповсюджуючись по дереву зверху вниз, починаючи з кореня і закінчуючи листовими елементами, відповідно до результатів тестів на шляху. На рисунку 1.1 зображено дерево прийняття рішень, яке прогнозує, чи буде клієнт підписуватися на електронну розсилку, чи ні (побудоване відповідно до наявної статистики). Варто зазначити, що дерева прийняття рішень здатні працювати як з дискретними, так і неперервними (числовими) атрибутами. Завдяки такому класифікатору, аналітик може спрогнозувати відклик потенційного клієнта щодо підписки на електронну розсилку. Кожен вузел підписано у відповідності до атрибуту, який перевіряється, а гілки — значеннями, відповідно до яких обирається шлях.

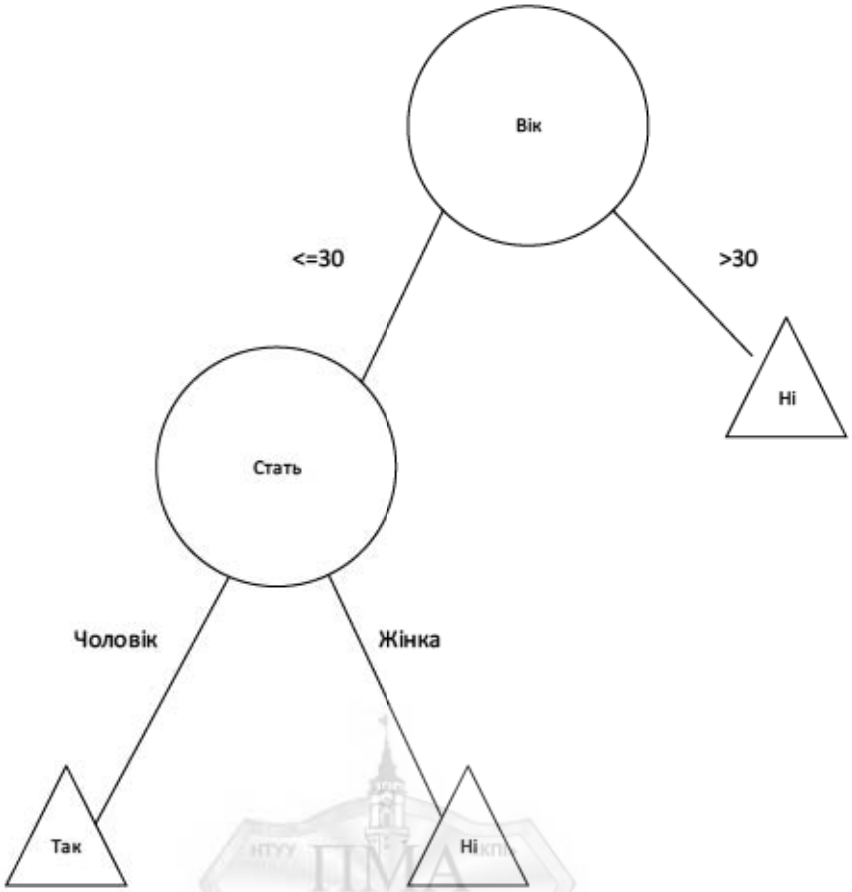

Рисунок 1.1 — Дерево прийняття рішень, яке прогнозує, чи буде клієнт підписуватися на електронну розсилку

У випадку чисельних атрибутів дерева прийняття рішень можуть бути інтерпретовані геометрично як набір гіперплощин, ортогональних до однієї з вісей. Природньо, що аналітики віддають перевагам менш складним деревам прийняття рівшень, оскільки їх простіше зрозуміти.

Складність дерева має значний вплив на точність його прогнозів. Вона визначається явно в залежності від критерію зупинки і методу обрізання дерева. Зазвичай, складність дерева вимірюються в одній з наступних метрик:

– загальна кількість вузлів;
– кількість листових вузлів;
– висота дерева;
– кількість атрибутів, використаних для побудови дерева.

Кожен шлях від кореня дерева до одного з його листів може бути трансформований у правило шляхом додавання частки «І» до тестів на шляху. Наприклад, один зі шляхів рисунка 1.1, може бути трансформовано у правило «Якщо вік клієнта менше 30 років і він чоловік, то він підпишеться на розсилку».

Індуктори дерев прийняття рішень — це алгоритми, які автоматично конструюють дерево прийняття рішень з відібраного набору даних. Зазвичай, ціллю являється пошук оптимального дерева прийняття рішень за критерієм мінімізації похибки узагальнення. Однак, можуть бути введено і додаткові критерії, наприклад, мінімізація висоти дерева, кількості вузлів, тощо.

Формування оптимального дерева прийняття рішень на основі наявних даних являється складною задачею. В [3] було доведено, що задача пошуку мінімального дерева, цілісного по відношенню до наявного тренувального набору, є NP-складною. Більш того, було показано, що конструювання мінімального двійкового дерева, яке здатне класифікувати очікувану кількість тестових прикладів, не знаючи їх заздалегіть, є NP-повною задачею [4].

Методи конструювання дерев прийняття рішень діляться на два класи в залежності від принципу конструювання дерева:

а) конструювання зверху-вниз;

б) конструювання знизу-вгору.

Переважна більшість відомих алгоритмів конструювання дерев прийняття рішень працює зверху-вниз. Серед них найбільшою популярністю користуються наступні:

а) ID3;

б) C4.5;

в) CART;

г) CHAID.

Такі алгоритми є «жадібними» і конструюють дерево зверху вниз за допомогою рекурсивних процедур (метод, також відомий як «розділяй і володарюй»). На кожній ітерації, алгоритм покриває частину тренувального набору в залежності від результуючого значення дискретної функції його атрибутів. Вибір найбільш

підходящої функції відбувається відповідно до визначеного способу розбиття. Після вибору підходящого зрізу, кожен вузол далі розділює тренувальний набір на менші піднабори, поки не залишиться пустий набір, або не буде задоволено критерій зупинки.

Для покращення прогнозів, які можна зробити за допомогою дерев прийняття рішень, використовують наступні підходи:

а) Bagging;

б) Random Forest;

в) Boosting.

Ці методи ґрунтуються на ідеї використання ансамблю дерев для побудови прогнозу.

Перевагами дерев прийняття рішень є те, що вони:

– можуть використовуватись для великих наборів даних;

– можуть використовуватись як для класифікації, так і для регресії;

– ігнорують незначущі для класифікації змінні;

– можуть працювати, навіть коли є пропуски у вибірці;

– стійкі до викидів у навчальній вибірці;

– малі дерева можна легко інтерпретувати.

Недоліками дерев прийняття рішень є те, що:

– великі дерева складно інтерпретувати;

– незначна зміна даних у навчальній вибірці може призвести до значної зміни структури дерева;

– немає механізму донавчання;

– дерева прийняття рішень без застосування додаткових методів покращення прогнозів, зазвичай, не можуть дати задовільну точність прогнозування.

1.2 Алгоритм ID3

ID3 є дуже простим алгоритмом дерев прийняття рішення [5]. Алгоритм ID3 використовує критерій приросту інформації як критерій розбиття. Критерій приросту інформації формулюється так:

$$InformationGain(a_i, S) =$$

$$= Entropy(y, S) - \sum_{v_{i,j} \in dom(a_i)} \frac{\left|\sigma_{a_i = v_{i,j}} S\right|}{|S|} \cdot Entropy\left(y, \sigma_{a_i = v_{i,j}} S\right),$$

де $Entropy(y, S) = \sum_{c_j \in dom(y)} -\frac{\left|\sigma_{y=c_j} S\right|}{|S|} \cdot \log_2 \frac{\left|\sigma_{y=c_j} S\right|}{|S|}$,

a_i — i-ий атрибут,

$v_{i,j}$ — значення i-того атрибута,

y — цільовий атрибут,

S — тренувальний набір.

Ріст дерева припиняється, коли всі екземпляри належать одному значенню цільового атрибуту, або коли найкращий приріст в інформації не більше 0. ID3 не використовує ніякі процедури обрізання дерев і не може працювати з чисельними атрибутами та даними, де є пропущені значення.

1.3 Алгоритм C4.5

C4.5 є алгоритмом, який базується на алгоритмі ID3, і запропонований тим же автором [6]. Він використовує критерій приросту як критерій розбиття. Критерій приросту формулюється так:

$$GainRation(a_i, S) = \frac{InformationGain(a_i, S)}{Entropy(a_i, S)},$$

де a_i — i-ий атрибут,

S — тренувальний набір.

Розбиття припиняється, коли кількість екземплярів розбиття нижче певного порогового значення. Засноване на похибці обрізання дерева виконується після фази росту. C4.5 може працювати з числовими атрибутами. Він може створювати дерево з тренувального набору, який містить відсутні значення за допомогою скоригованого критерію приросту.

1.4 Алгоритм CART

Алгоритм CART розшифровується як Classification and Regression Trees (дерева класифікації та регресії) [7]. Він характеризується тим, що конструює бінарні дерева, іншими словами, кожен внутрішній вузол має рівно дві дуги-виходи. Розбиття виконується у відповідності до критерію буксирування. Критерій буксирування формулюється так:

$$twoing(a_i, dom_1(a_i), dom_2(a_i), S) = 0{,}25 \cdot \frac{\left|\sigma_{a_i \in dom_1(a_i)} S\right|}{|S|} \cdot \frac{\left|\sigma_{a_i \in dom_2(a_i)} S\right|}{|S|} \cdot \left(\sum_{c_i \in dom(y)} \left| \frac{\left|\sigma_{a_i \in dom_1(a_i)\ AND\ y=c_i} S\right|}{\left|\sigma_{a_i \in dom_1(a_i)} S\right|} - \frac{\left|\sigma_{a_i \in dom_2(a_i)\ AND\ y=c_i} S\right|}{\left|\sigma_{a_i \in dom_2(a_i)} S\right|} \right| \right)^2$$

де a_i — i-ий атрибут,

$dom_j(a_i)$ — j-тий піддомен i-ого атрибуту,

$v_{i,j}$ — значення i-того атрибута,

y — цільовий атрибут,

S — тренувальний набір.

Отримане дерево обрізається за допомогою методу обрізання вартість-складність. За необхідності CART може враховувати вартість неправильної класифікації при побудові дерева. Також він дозволяє користувачам отримати апріорний розподіл ймовірностей.

Важливою перевагою CART є здатність методу створювати дерева регресії. Дерева регресії — це такі дерева прийняття рішень, листові вузли яких прогнозують дійсні числа, а не клас. У випадку регресії, CART шукає розбиття, які мінімізують середньоквадратичну похибку. Прогноз на кожному листі базується на зваженому середньому значень вузлів.

1.5 Алгоритм CHAID

Починаючи з 70-х років попереднього сторіччя, вчені, які займаються прикладною статистикою, розроблюють процедури генерації дерев прийняття рішень, такі як: AID [8], MAID [9], THAID [10], CHAID [11]. CHAID (Chisquare–Automatic–Interaction–Detection) був спочатку розроблений для того, щоб працювати лише з категоріальними змінними. Для кожного вхідного атрибута a_i, CHAID знаходить пари значень, які найменш значимо відрізняються від цільової змінної. Відмінність вимірюється за допомогою p-value, отриманого зі статистичного тесту. Статистичний тест відрізняється в залежності від того, який тип має цільовий атрибут (неперервний, впорядкований чи категоріальний).

Для кожної відібраної пари, CHAID перевіряє чи p-value більше за певне порогове значення. Якщо так, то воно об’єднує значення і шукає додаткову

потенційну пару для об'єднання. Процес повторюється до тих пір, поки не будуть знайдені всі значимі пари.

Після цього обирається найкращий вхідний атрибут для розбиття в поточному вузлі. Якщо p-value найкращого атрибута більше певного порогового значення, то розбиття не відбувається. Також ця процедура зупиняється, коли одна з наступних умов виконана:

а) досягнута максимальна глубина дерева;

б) мінімальна кількість екземплярів у вузлі досягнута, так що не можна розділювати простір далі.

CHAID не виконує процедуру обрізання дерева. Він добре підходить для дослідження предметної області, але такі методи як CART та C4.5 перевершують його у випадку вирішення задачі прогнозування.

1.6 Метод Bagging

Вперше метод був запропонований у 1994 році в [12]. Він полягає у генерації декількох версій предиктора і застосуванні їх для отримання агрегованого предиктора. Агрегація усереднює значення кожного предиктора для задач регресії і використовує процедуру голосування (вибір зупиняється на результаті, за який більшість голосів) для задач класифікації. Декілька версій предиктора формуються за рахунок відбору підвибірок з вихідної навчальної вибірки і використання їх як навчальних виборок для предикторів. Тести на реальних і модельних даних показають, що в результаті використання методу bagging вдається досягти значного підвищення точності результатів. Важливим є елемент нестабільності в методі прогнозування — в такому випадку процедура bagging'у може значно покращити точність.

1.7 Сімейство методів Random Forest

Випадкові ліси — сімейство методів, які грунтуються на побудові ансамблю (або лісу) дерев прийняття рішень, які вирощуються з рандомізованої варіації алгоритму вирощування дерева [13]. Дерева прийняття рішень найбільш підходять для використанння підходу на основі ансамблю предикторів, оскільки вони мають низьке зміщення і високу варіативність, що надає їм переваги від процесу усереднення.

Алгоритм формування випадкового лісу наступний [14]:

I) повторити з $b = 1$ по B:

а) обрати випадковим чином вибірку Z^* розміру N з навчальних даних.

б) виростити дерево T_b для випадкового лісу на основі вибірки Z^*, шляхом рекурсивого повторнення наступних кроків для кожного термінального вузла дерева, поки не досягнуто мінімальний розмір вузла n_{min}:

1) обрати m змінних випадковим чином з p змінних;

2) обрати найкразу змінну для розбиття серед цих m змінних;

3) побудувати розбиття.

II) зберегти ансамбль дерев $\{T_b\}_1^B$.

Для того, щоб отримати прогноз в новій точці x, використовуються спеціальні формули. Якщо маємо справу з задачею регресії:

$$\hat{f}_{rf}^B(x) = \frac{1}{B}\sum_{b=1}^{B} T_b(x),$$

Якщо ж з задачею класифікації, то:

$$\hat{C}_{rf}^B(x) = majority\ vote\ \{\hat{C}_b(x)\}_1^B,$$

де $\hat{C}_b(x)$ — результат прогнозування дерева під номером b,

$majority\ vote$ — функція, яка повертає той клас, за який проголосувало найбільша кількість дерев.

Фактично, цей метод є покращенням методу bagging (додано вибір m випадкових змінних при роботі з вибірками).

1.8 Метод Boosting

Метод полягає у використанні покращеного класифікатора (або регресора) над слабкими класифікаторами (регресорами) [15]. Покращений класифікатор має наступний вигляд:

$$H(\bar{x}) = \sum_t \alpha_t h_t(\bar{x}),$$

де $H(\bar{x})$ — результат прогнозування для набору атрибутів $\bar{x}$,

t — кількість слабких класифікаторів,

α_t — вага впливу t-ого класифікатора на результат,

$h_t(\bar{x})$ — t-ий слабкий класифікатор.

Ваги α_t підбираються таким чином, щоб покращений класифікатор мав якомого вищу точність прогнозування. Існує декілька аглоритмів для навчання класифікатора. Найбільш популярним є алгоритм AdaBoost (Adaptive Boosting) [16]. Він розраховує ваги α_t за наступною формулою:

$$\alpha_t = \beta \cdot \ln\left(\frac{1 - err_t}{err_t}\right),$$

де β — нормалізуюча константа,

err_t — похибка ваги, яка розраховується як:

$$err_t = \sum_i W_{i,t} e^{-y_i f(x_i)},$$

де $W_{i,t}$ — вага,

$h_t(\bar{x})$ — t-ий слабкий класифікатор,

y — бажане результуюче значення.

Класифікатори $h_t(\bar{x})$ мають навчатися таким чином, щоб мінімізувати похибку err_t. Згідно з [17], Boosting має кращу ефективність у порівнянні з підходами на основі Random Forest, а Random Forest має кращу ефективність ніж Bagging (тобто, у перспективі може дати більший приріст точності прогнозування). Як зрозуміло з опису методу, його можна вважати розширенням Random Forest, в якому результату кожного з навчених дерев лісу приписано певну вагу.

1.9 Висновки до розділу

Дерево прийняття рішень є класифікатором, який виконує рекурсивне розбиття простору екземплярів. Перевагами дерев прийняття рішень є те, що вони:

- можуть використовуватись для великих наборів даних;
- можуть використовуватись як для класифікації, так і регресії;
- ігнорують незначущі для класифікації змінні;
- можуть працювати навіть коли є пропуски у вибіріці;
- стійкі до викидів у навчальній вибірці;
- малі дерева можна легко інтерпретувати.

Точність прогнозів дерев прийняття рішень без використання додаткових методів покращення прогнозів, зазвичай, нижче, ніж у інших методів машинного

навчання (методу опорних векторів, нейронних мереж), але застосовуючи такі методи як Bagging, Random Forest чи Boosting можна іноді досягти результатів прогнозування кращих, ніж за рахунок використання інших методів прогнозування.

У дослідженні вирішено використати як метод C4.5, так і метод CART, оскільки складно заздалегіть сказати, який з цих методів спрацює краще. C4.5 і CART є більш популярними ніж CHAID і ID3. ID3 не має сенсу використовувати, оскільки метод C4.5 є його прямим нащадком і має краще показники. CHAID краще підходить для аналізу предметної області, ніж прогнозування, але оскільки в даній роботі вирішується задача прогнозування, то краще застосовувати методи CART і C4.5. Крім того, обрано такі методи як Random Forest і Boosting (з алгоритмом AdaBoost) на основі ансамблів дерев прийняття рішень.

2 РЕАЛІЗАЦІЯ МЕТОДУ ПРОГНОЗУВАННЯ НГ НА БАЗІ ДЕРЕВ ПРИЙНЯТТЯ РІШЕНЬ

Як було зазначено у підрозділі 1.9, для прогнозування випадків нічної гіпоглікемії обрано такі методи як C4.5, CART, Random Forest і AdaBoost. Всі ці методи є непараметричними і реалізовані в багатьох мовах програмування та математичних пакетах (наприклад, у R [18], Python [19] та інших). Головною задачею, яка постає перед дослідником, є підготовка даних для навчання і тестування моделей, а також вибір параметрів методів.

2.1 Особливості застосування методу на підготовлених даних з проекту DirecNet

Оскільки обрані моделі є непараметризованими, то для їх налаштування і використання з даними проекту DirecNet не потрібно виконувати додаткові дії. Таблиця даних, що використовується для навчання і тестування моделей, складається зі стовпців, наведених у таблиці 4.6 спільної частини, за виключенням стовпців NoctMin_Gl та NoctMin_T. Вказані стовпці було виключено з таблиці, щоб уникнути автокореляції зі значеннями стовпця Hypoglycemia.

Входами моделей в залежності від набору даних були:

– показники рівня глюкози в крові (стовпці BMax_Gl1, Max_Gl1, BMax_Gl2, Max_Gl2, BMax_Gl3, Max_Gl3, TheLastBeforeBed);

– часи замірів рівня глюкози (стовпці BMax_T1, Max_T1, BMax_T2, Max_T2, BMax_T3, Max_T3);

– швидкості росту рівня глюкози (стовпці V1, V2, V3);

– демографічні дані (стовпці Ill_years, Age, Gender, InsMod);

– додаткові фізіологічні дані (стовпці Height, Weight, HbA1C, BMI).

Виходами моделей вважалися категоріальні значення стовпця Hypoglycemia.

Підбір конкретних параметрів моделей відбувається автоматизовано за допомогою засобів статистичної системи R. Як критерій оптимальності використовується точність прогнозування, тобто кількість правильно ідентифікованих випадків гіпоглікемії (та її відсутності).

Для алгоритму CART підбиралися наступні параметри:

– кількість внутрішніх вузлів;

– мінімальна кількість екземплярів, які мають бути у підпросторі, для побудови вузла;

– мінімальна кількість екземплярів у листовому вузлі.

Для алгоритму C4.5:

– кількість проходів для вибору оптимального по критерію точності дерева;

– мінімальна кількість екземплярів у листовому вузлі;

– значення довірчого порогу для обрізання дерева.

Для методу Random Forest:

– кількість дерев у лісі;

– значення змінної m (кількість випадково обираємих змінних);

– розмір вибірки, обраної за допомогою процедури бутстрапінгу;

– мінімальна кількість екземплярів у листовому вузлі дерева;

– максимальна кількість листових вузлів у дереві.

Для методу AdaBoost:

– кількість дерев для навчання;

– кількість внутрішніх вузлів;

– мінімальна кількість екземплярів, які мають бути у підпросторі, для побудови вузла;

– мінімальна кількість екземплярів у листовому вузлі.

2.2 Програмна реалізація методу

Як мову програмування для реалізації методів було вирішено використати R [20]. R є системою для статистичних обрахунків і візуалізації. Вона складається з одноіменної мови програмування і середовища виконання з графічним інтерфейсом і відладчиком. Основою R є інтерпретована комп'ютерна мова, яка дозволяє використовувати розвітвлення і циклічне виконання операцій та модульне програмування з функціями. Більшість доступних користувачу функцій написана засобами самої ж мови. Крім того, доступна інтеграція R у інші мови програмування, такі як C, C++ і FORTRAN. R включає в себе великий набір статистичних функцій для обробки масивів даних. Серед них: лінійна і узагальнена лінійна регресії, нелінійні регресійні моделі, функції для аналізу часових рядів, класичні параметричні і непараметричні тести, кластеризація і згладжування. З метою розширення функціональності можна використовувати додаткові модулі.

У програмній реалізації даного методу прогнозування було використано додаткові бібліотеки, такі як:

– rpart — бібліотека для конструювання дерев за допомогою алгоритма CART;

– RWeka — бібліотека для конструювання дерев за допомогою алгоритма C4.5;

– randomForest — бібліотека для конструювання випадкових лісів (метод Random Forest) на основі дерев CART;

– ada — бібліотека для використання методу Boosting на основі алгоритму AdaBoost над деревами CART;

– rpart.plot — бібліотека для побудови графіків дерев прийняття рішень;

– caret — бібліотека для факторного аналізу, побудови матриці помилок та розрахунку метрик помилок.

Графічний інтерфейс користувача було створена за допомогою засобів мови програмування C++ і бібліотеки wxWidgets.

Опис модулів, з яких складається програмна реалізація методу прогнозування на основі дерев прийняття рішень, наведено в таблиці 2.1. Зовнішній вигляд екранних форм приведено на рисунках 2.1 та 2.2. Схема алгоритму роботи програми наведена на рисунку 2.3.

Таблиця 2.1 — Модулі програмної реалізації методу прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу

Модуль	Опис
config.r	Налаштування шляхів і підключення необхідних бібліотек для коректної роботи модулів, написаних на R
fixed_dataset_v4_boosting.r	Модуль, який використовує метод AdaBoost на основі дерев CART для прогнозування нічної гіпоглікемії
fixed_dataset_v4_c4_5.r	Модуль, який використовує метод C4.5 для прогнозування нічної гіпоглікемії
fixed_dataset_v4_cart.r	Модуль, який використовує метод CART для прогнозування нічної гіпоглікемії
fixed_dataset_v4_random_forest.r	Модуль, який використовує метод Random Forest на основі дерев CART для прогнозування нічної гіпоглікемії
metrics.r	Функції для розрахунку значень метрик з матриці помилок (таких як TPR, FPR, NPV, PPV, точність прогнозу, F1, MCC), а також запису отриманих значень в файл формату CSV
prepare-test-datasets.r	Функції для виокремлення наборів даних з одного загального набору

Продовження таблиці 2.1.

Модуль	Опис
Application.h	Заголовок, який описує головний клас програми
DialogAbout.h	Заголовок, який описує діалогове вікно «Про авторів»
WindowMain.h	Заголовок, який описує головне вікно програми
Application.cpp	Реалізація головного класу програми
DialogAbout.cpp	Реалізація діалогового вікна «Про авторів»
WindowMain.cpp	Реалізація головного вікна програми

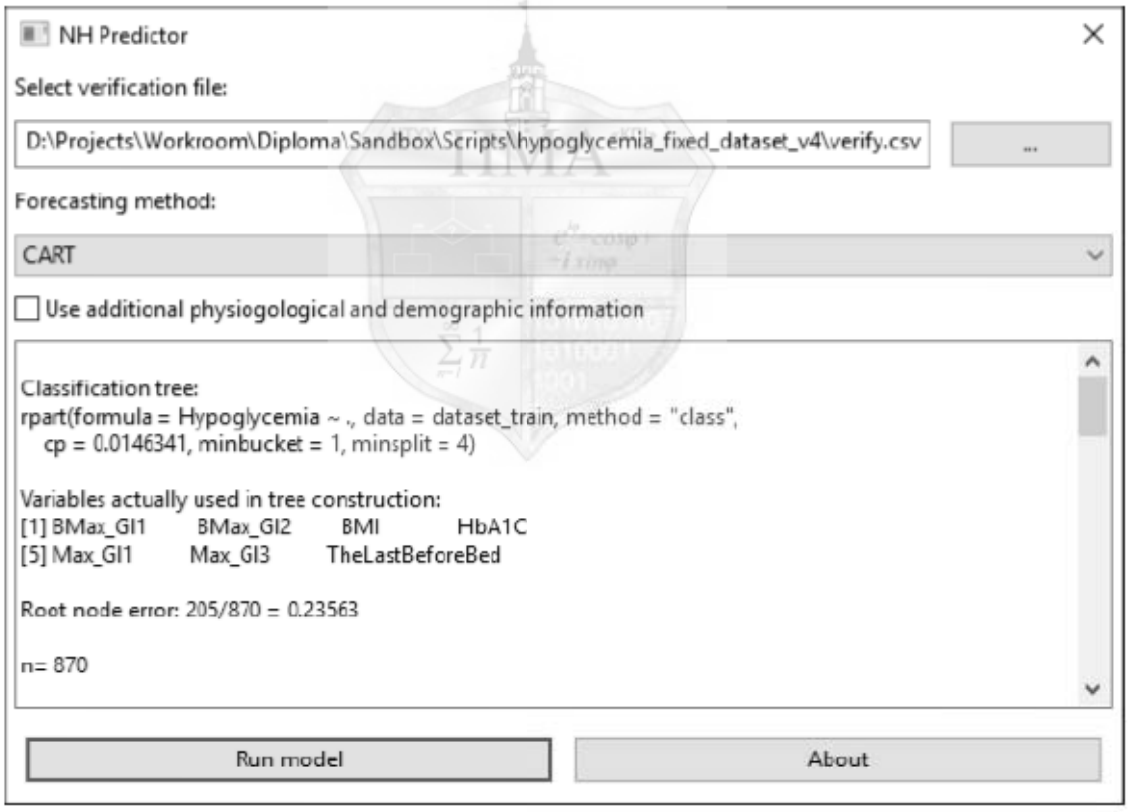

Рисунок 2.1 — Головне вікно програми

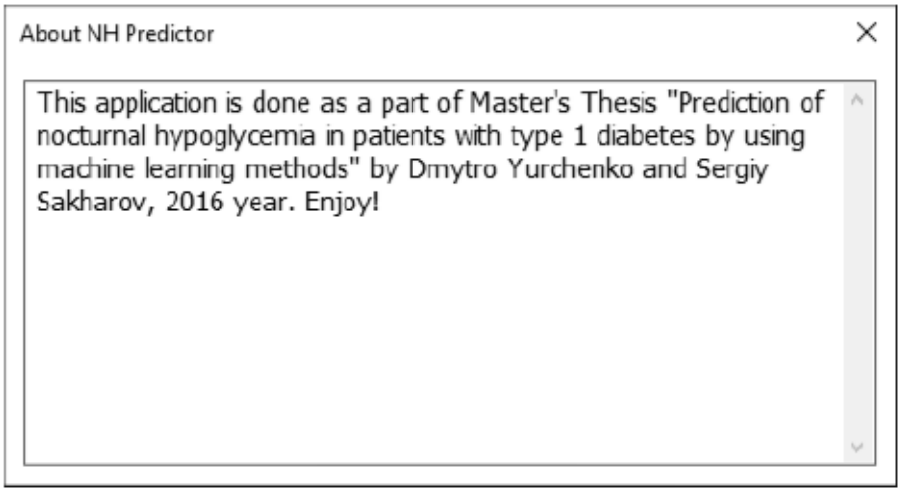

Рисунок 2.2 — Діалогове вікно «Про авторів»

Рисунок 2.3 — Схема алгоритму роботи програми

Програма приймає на вхід .CSV файл із записами пацієнтів і тестує роботу обраного методу на них. Значенння метрик похибок і значимості факторів для побудови прогнозу виводиться на екранну форму, що дозволяє порівняти роботу різних методів прогнозування.

2.3 Висновки до розділу

Методи прогнозування було пристосовано до роботи з даними DirecNet. Підбір конкретних параметрів моделей відбувається автоматизовано за допомогою засобів статистичної системи R. Як критерій оптимальності використовується точність прогнозування, тобто кількість правильно ідентифікованих випадків гіпоглікемії (та її відсутності).

Для тестування методів розроблено програмне забезпечення, яке приймає на вхід CSV-файл із записами пацієнтів і тестує роботу обраного методу. Програмне забезпечення виконане засобами мови програмування R з застосуванням додаткових бібліотек rpart, randomForest, RWeka, ada, rpart.plot, caret та C++ з застосуванням бібліотеки wxWidgets.

3 РЕЗУЛЬТАТИ ЕКСПЕРИМЕНТАЛЬНОГО ДОСЛІДЖЕННЯ

Після того, як було обрано методи прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу, і розроблено програмне забезпечення, що їх реалізує, проводиться тестування методів на даних. Для порівняння методів мають бути обрані метрики, які дадуть можливість адекватно оцінити ефективність роботи методів. Після цього відбувається тестування методів на заздалегіть підготовлених наборах даних, які включають всі наявні атрибути або лише частину з них. Окрім того, цікавою постає задача порівняння впливу різних факторів на якість прогнозування, що дозволить у подальшому вдосконалити наявні методи прогнозування нічної гіпоглікемії.

3.1 Метрики порівняння результатів роботи методів

Для порівняння результатів роботи моделей, які отримані в результаті застосування алгоритмів CART, C4.5 та методів Random Forest та AdaBoost, було використано класичні метрики порівняння двійкових класифікаторів, засновані на матриці похибок (див. таблицю 3.1) [21].

Таблиця 3.1 — Приклад матриці похибок

	Прогнозване «Так»	Прогнозоване «Ні»
Справжнє «Так»	$TP = 100$	$FN = 20$
Справжнє «Ні»	$FP = 30$	$TN = 80$

Матриця похибок представляє собою таблицю значень 2x2, яка містить наступні цілочисельні значення:

— TP (True Positive) — кількість прогнозованих значень «Так», які співпали зі справжніми значеннями «Так» (вказаними у тестовій вибірці);

— FN (False Negative) — кількість прогнозованих значень «Ні», які видані, коли справжніми значеннями були «Так» (помилка 2-го роду);

— FP (False Positive) — кількість прогнозованих значень «Так», які видані, коли справжніми значеннями були «Ні» (помилка 1-го роду);

— TN (True Negative) — кількість прогнозованих значень «Ні», які співпали зі справжніми значеннями «Ні».

На основі цих значень розраховуються значення наступних метрик:

$$TPR = \frac{TP}{TP + FN},$$

$$TNR = \frac{TN}{TN + FP},$$

$$PPV = \frac{TP}{TP + FP},$$

$$NPV = \frac{TN}{TN + FN},$$

$$ACC = \frac{TP + TN}{TP + FP + FN + TN},$$

$$F1 = \frac{2TP}{2TP + FP + FN},$$

$$MCC = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}},$$

де TPR — True Positive Rate, частота позитивних прогнозів,

TNR — True Negative Rate, частота негативних прогнозів,

PPV — Positive Predictive Value, значимість позитивних прогнозів,

NPV — Negative Predictive Value, значимість негативних прогнозів,

ACC — Accuracy, точність прогнозування,

$F1$ — F1 score, гармонічне середнє PPV і TPR,

MCC — Matthews correlation coefficient, коефіцієнт кореляції Метью.

TPR вказує ймовірність того, що буде даний позитивний прогноз пацієнту, у якого буде приступ гіпоглікемії. Варіюється від 0 до 1, найкращим значенням є 1.

TNR вказує ймовірність того, що буде даний негативний прогноз пацієнту, у якого не буде приступу гіпоглікемії. Варіюється від 0 до 1, найкращим значенням є 1.

PPV вказує, який відсоток позитивних прогнозів, які будуть дані, справдиться. Варіюється від 0 до 1, найкращим значенням є 1.

NPV вказує, який відсоток негативних прогнозів, які будуть дані, справдиться. Варіюється від 0 до 1, найкращим значенням є 1.

ACC вказує, який відсоток прогнозів буде відповідати дійсності. Варіюється від 0 до 1, найкращим значенням є 1.

$F1$ — міра того, наскільки добре предиктор здатен робити прогнози позитивних значень. Варіюється від 0 до 1, найкращим значенням є 1.

MCC — міра того, наскільки добре предиктор здатен робити прогнози як позитивних, так і негативних значень. Варіюється від -1 до 1, найкращим значенням є 1. Значенню 0 відповідає здатність прогнозувати не краще, ніж просто випадковим чином. Значенню -1 відповідає повна нездатність моделі прогнозувати (ще гірше, ніж прогнозування випадковим чином).

3.2 Набори даних для тестування методів

Для застосування методів прогнозування було підготовлено 17 наборів даних з загального набору даних DirecNet, описаного у підрозділі 4.4 спільної частини. Ці набори відрізняються тим, які атрибути пацієнта вони містять. На навчальну вибірку було відведено 80% екземплярів набору даних (870 діб), на тестувальну — 20% (218 діб). Розподіл даних між навчальною та тестувальною вибіркою виконано випадковим чином. У таблиці 3.2 наведено перелік наборів даних і їх відмінності.

Таблиця 3.2 — Підготовані набори даних

№ набору	Рівні глюкози	Час замірів	Швидкості росту	Демографічні дані
1	+	-	-	-
2	+	+	-	-
3	+	-	+	-
4	+	+	+	-
5	+	-	-	Ill_years, Age, Gender, Height, Weight, InsMod, HbA1C, BMI
6	+	+	-	Ill_years, Age, Gender, Height, Weight, InsMod, HbA1C, BMI
7	+	-	+	Ill_years, Age, Gender, Height, Weight, InsMod, HbA1C, BMI
8	+	+	+	Ill_years, Age, Gender, Height, Weight, InsMod, HbA1C, BMI
9	+	+	+	Ill_years, Age, Gender, InsMod, HbA1C, BMI
10	+	+	+	Ill_years
11	+	+	+	Age
12	+	+	+	Gender
13	+	+	+	Height
14	+	+	+	Weight
15	+	+	+	InsMod
16	+	+	+	HbA1C
17	+	+	+	BMI

3.3 Результати роботи моделей

Роботу моделей було протестовано на 17 наборах даних, описаних у підрозділі 3.2. Отримані результати для алгоритму CART знаходяться у таблиці 3.3, для алгоритму C4.5 — у таблиці 3.4, для методу Random Forest — у таблиці 3.5, для методу AdaBoost — у таблиці 3.6.

Також для порівняння було отримано результати роботи методу на основі лінійної комбінації предикторів [22]. Порівняння роботи розроблених методів і наявних наведено у таблиці 3.7.

Таблиця 3.3 — Результати роботи алгоритму CART

Набір даних	TPR	TNR	PPV	NPV	ACC	F1	MCC
1	0,37	0,81	0,34	0,83	0,72	0,35	0,17
2	0,57	0,79	0,42	0,87	0,74	0,48	0,32
3	0,35	0,8	0,32	0,82	0,71	0,33	0,15
4	0,57	0,78	0,41	0,87	0,74	0,48	0,32
5	0,37	0,83	0,36	0,83	0,73	0,37	0,19
6	0,48	0,76	0,35	0,85	0,7	0,4	0,22
7	0,41	0,81	0,37	0,84	0,72	0,39	0,21
8	0,41	0,78	0,34	0,83	0,71	0,37	0,18
9	0,41	0,78	0,33	0,83	0,7	0,37	0,18
10	0,54	0,78	0,4	0,86	0,73	0,46	0,29
11	0,57	0,78	0,41	0,87	0,74	0,48	0,32
12	0,57	0,78	0,41	0,87	0,74	0,48	0,32
13	0,52	0,79	0,4	0,86	0,73	0,45	0,29
14	0,52	0,8	0,41	0,86	0,74	0,46	0,29
15	0,57	0,78	0,41	0,87	0,74	0,48	0,32
16	0,43	0,78	0,35	0,84	0,71	0,39	0,2
17	0,57	0,81	0,44	0,87	0,76	0,5	0,34

Таблиця 3.4 — Результати роботи алгоритму C4.5

Набір даних	TPR	TNR	PPV	NPV	ACC	F1	MCC
1	0,17	0,97	0,57	0,81	0,8	0,27	0,23
2	0,2	0,92	0,39	0,81	0,77	0,26	0,15
3	0,17	0,97	0,57	0,81	0,8	0,27	0,23
4	0,2	0,95	0,5	0,82	0,79	0,28	0,21
5	0,28	0,87	0,36	0,82	0,74	0,32	0,16
6	0,46	0,79	0,37	0,84	0,72	0,41	0,23
7	0,24	0,88	0,34	0,81	0,74	0,28	0,13
8	0,37	0,8	0,33	0,83	0,71	0,35	0,16
9	0,41	0,77	0,33	0,83	0,7	0,37	0,17
10	0,37	0,83	0,37	0,83	0,73	0,37	0,2
11	0,2	0,94	0,45	0,81	0,78	0,27	0,19
12	0,39	0,85	0,42	0,84	0,76	0,4	0,25
13	0,2	0,94	0,45	0,81	0,78	0,27	0,19
14	0,2	0,95	0,5	0,82	0,79	0,28	0,21
15	0,24	0,91	0,41	0,82	0,77	0,3	0,18
16	0,22	0,92	0,43	0,82	0,78	0,29	0,19
17	0,3	0,93	0,54	0,83	0,8	0,39	0,3

Таблиця 3.5 — Результати роботи методу Random Forest

Набір даних	TPR	TNR	PPV	NPV	ACC	F1	MCC
1	0,2	0,95	0,5	0,82	0,79	0,28	0,21
2	0,3	0,97	0,7	0,84	0,83	0,42	0,38
3	0,22	0,94	0,5	0,82	0,79	0,3	0,23
4	0,33	0,95	0,63	0,84	0,82	0,43	0,36
5	0,39	0,91	0,55	0,85	0,8	0,46	0,35
6	0,3	0,95	0,64	0,84	0,82	0,41	0,35
7	0,3	0,93	0,54	0,83	0,8	0,39	0,3
8	0,28	0,96	0,65	0,83	0,82	0,39	0,34
9	0,3	0,96	0,67	0,84	0,82	0,42	0,36
10	0,28	0,94	0,57	0,83	0,8	0,38	0,3
11	0,3	0,95	0,64	0,84	0,82	0,41	0,35
12	0,28	0,97	0,68	0,83	0,82	0,4	0,36
13	0,3	0,94	0,56	0,83	0,8	0,39	0,31
14	0,33	0,94	0,6	0,84	0,81	0,42	0,34
15	0,28	0,95	0,59	0,83	0,81	0,38	0,31
16	0,3	0,95	0,61	0,84	0,81	0,41	0,33
17	0,33	0,95	0,63	0,84	0,82	0,43	0,36

Таблиця 3.6 — Результати роботи методу AdaBoost

Набір даних	TPR	TNR	PPV	NPV	ACC	F1	MCC
1	0,26	0,94	0,55	0,83	0,8	0,35	0,27
2	0,28	0,96	0,65	0,83	0,82	0,39	0,34
3	0,3	0,92	0,52	0,83	0,79	0,38	0,28
4	0,39	0,91	0,55	0,85	0,8	0,46	0,35
5	0,41	0,92	0,59	0,85	0,82	0,49	0,39
6	0,43	0,91	0,57	0,86	0,81	0,49	0,39
7	0,41	0,93	0,61	0,86	0,82	0,49	0,4
8	0,39	0,95	0,69	0,85	0,83	0,5	0,43
9	0,37	0,94	0,61	0,85	0,82	0,46	0,37
10	0,3	0,91	0,47	0,83	0,78	0,37	0,25
11	0,37	0,94	0,61	0,85	0,82	0,46	0,37
12	0,35	0,94	0,62	0,84	0,82	0,44	0,36
13	0,37	0,92	0,57	0,85	0,81	0,45	0,35
14	0,35	0,92	0,55	0,84	0,8	0,43	0,33
15	0,39	0,92	0,58	0,85	0,81	0,47	0,37
16	0,35	0,94	0,59	0,84	0,81	0,44	0,35
17	0,41	0,91	0,54	0,85	0,8	0,47	0,36

Таблиця 3.7 — Порівняльна таблиця результатів роботи методів на основі лінійної комбінації предикторів і розроблених методів

Існуючі методи							
Метод	TPR	TNR	PPV	NPV	ACC	F1	MCC
DIAppvisor_risk	0,44	0,85	0,46	0,83	0,75	0,45	0,29
DIAppvisor	0,74	0,55	0,33	0,87	0,59	0,45	0,24
Розроблені методи							
Метод	TPR	TNR	PPV	NPV	ACC	F1	MCC
CART	0,57	0,81	0,44	0,87	0,76	0,5	0,34
C4.5	0,3	0,93	0,54	0,83	0,8	0,39	0,3
Random Forest	0,3	0,96	0,67	0,84	0,82	0,42	0,36
AdaBoost	0,39	0,95	0,69	0,85	0,83	0,5	0,43

Проаналізуємо таблиці з отриманими результатами. Найкращі результати алгоритму CART ($MCC = 0{,}34$) були отримані на 17-ому наборі даних. Параметри методу наступні:

— кількість внутрішніх вузлів дорівнює 85;

— мінімальна кількість екземплярів, які мають бути у підпросторі, для побудови вузла дорівнює 5;

— мінімальна кількість екземплярів у листовому вузлі дорівнює 2.

Значення $F1 = 0{,}5$ і воно є найвищим серед всіх моделей, що каже про хорошу здатність моделі ідентифікувати випадки, коли гіпоглікемія матиме місце. Не дуже очевидним є те, що метод так добре спрацював саме на 17-ому наборі, а не на, скажімо, 8-ому, де є всі ці атрибути і більше. З цього можна зробити висновок, що є певна зашумленість у значеннях інших атрибутів, яка не дає можливість методу розкрити себе найкращим чином.

Результати роботи алгоритмів CART та C4.5 показали, що наявність зросту і ваги в даних не лише не впливає на якість прогнозування, а іноді навіть погіршує (це можна побачити на результатах методів у таблицях 3.3-3.6, на наборах даних 8, 13 та 14). Швидше за все, це пов'язано з тим, що дані зросту і ваги пацієнтів не мають

однозначного зв'язку з ймовірністю випадків гіпоглікемії (див. рисунки 3.17 та 3.18 спільної частини) і вносять шум. Методи Random Forest і AdaBoost, тим не менше змогли використати ці атрибути для покращення результату прогнозування.

Найкращі результати алгоритму C4.5 ($MCC = 0{,}3$) також були отримані на 17-ому наборі даних. Метод проявив себе гірше за показниками MCC і $F1$, ніж CART та краще по параметрам ACC, PPV і TNR. Але хоч TNR і вище, значення NPV нижче, з чого можна зробити висновок, що прогнозам моделі треба довіряти менше. Параметри методу наступні:

— кількість проходів для вибору оптимального по критерію точності дерева дорівнює 3;

— мінімальна кількість екземплярів у листовому вузлі дорівнює 2;

— значення довірчого порогу для обрізання дерева дорівнює 0,25.

Найкращі результати методу Random Forest ($MCC = 0{,}36$) були отримані на 9-ому та 17-ому наборах даних. Параметри методу наступні:

— кількість дерев у лісі дорівнює 500;

— значення змінної m дорівнює 4 (кількість випадково обираємих змінних);

— розмір вибірки, обраної за допомогою процедури бутстрапінгу дорівнює 870;

— мінімальна кількість екземплярів у листовому вузлі дерева дорівнює 5;

— максимальна кількість листових вузлів у дереві необмежена.

Найкращі результати методу AdaBoost ($MCC = 0{,}43$) були отримані на 8-ому наборі даних. Цей метод перевершує усі інші методи по кожній з метрик. Параметри методу наступні:

— кількість дерев для навчання дорівнює 100;

— кількість внутрішніх вузлів у дереві дорівнює 435;

— мінімальна кількість екземплярів, які мають бути у підпросторі, для побудови вузла дорівнює 5;

— мінімальна кількість екземплярів у листовому вузлі дорівнює 2.

У процесі генерації моделей було виявлено, що кінцеві моделі, які генеруються за допомогою алгоритмів CART та C4.5 визначаються детерміновано за допомогою параметрів методу і можуть бути відтворені шляхом вибору таких самих параметрів і навчання на тих же даних. Random Forest і AdaBoost в силу того, що в алгоритм закладена випадковість вибору підвибірки допускають флуктуації параметрів під час тренування, тобто інколи можна отримати модель, показник MCC якої значно менший, ніж цього можна досягти. Тому під час генерації моделей методами Random Forest і AdaBoost відбиралися такі моделі, які показали найкращий результат по показнику MCC на тренувальній вибірці (для кожної вибірки генерувалося 100 моделей, а з них обиралася одна найкраща).

3.4 Аналіз впливу факторів на результат прогнозування

На основі найкращої моделі з отриманих (AdaBoost на 8-му наборі даних) було побудовано графік впливу факторів на результат прогнозування (рисунок 3.1). Демографічні фактори, такі як зріст, вік, вага, показник BMI вважаються важливішими навіть за конкретні значення замірів рівня глюкози.

Менш досконалі методи не змогли виявити такої значимості впливу факторів на прогноз. Наприклад, графік впливу факторів на результат прогнозування для моделі CART на 8-му наборі даних приведено на рисунку 3.2. Найбільш важливим фактором є значення глікемії перед сном. Фактори часу замірів також значно впливають на результат прогнозування. Всі інші фактори є менш значущими.

Зрозуміло, що правильний вибір факторів значно впливає на результати прогнозування і оскільки модель CART і C4.5 не вважають демографічні показники важливими для прогнозування, то отримані результати відповідно нижчі.

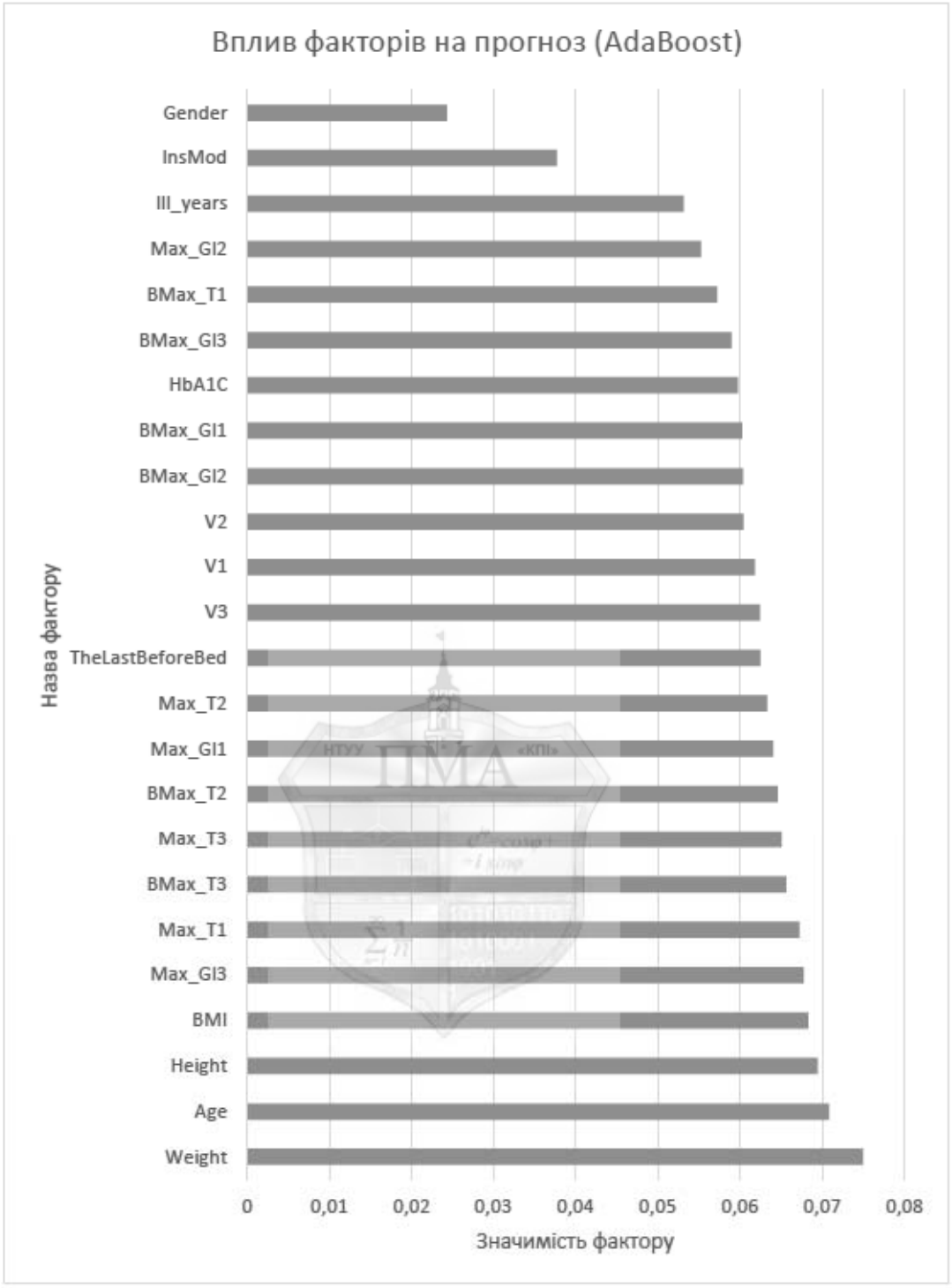

Рисунок 3.1 — Значимості факторів для побудови прогнозу нічної гіпоглікемії за допомогою отриманої методом AdaBoost моделі

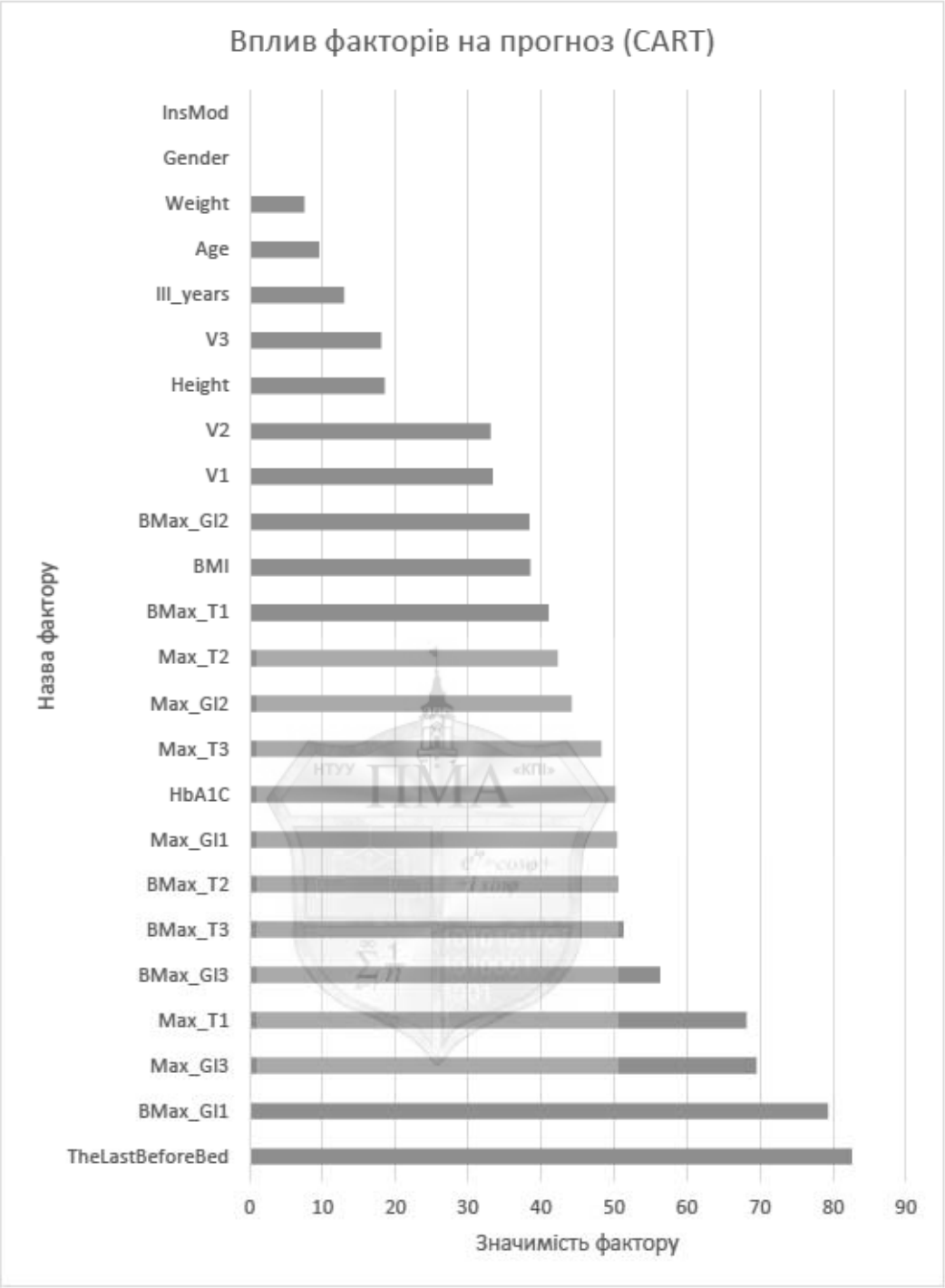

Рисунок 3.2 — Значимості факторів для побудови прогнозу нічної гіпоглікемії за допомогою отриманої алгоритмом CART моделі

Всі методи вважають фактори статі і схеми лікування неважливими для прогнозу.

3.5 Висновки до розділу

У розділі описано метрики, що використовуються для порівняння різних моделей прогнозування нічної гіпоглікемії у хворих на діабет 1-го типу. З вихідної вибірки даних DirecNet було сформовано 17 вибірок, які відрізнялися атрибутами. Роботу реалізованих методів було протестовано на кожній з цих вибірок і обрано найкращу модель, яку може дати метод.

Було порівняно результати прогнозування розроблених моделей з наявними рішеннями (див. таблицю 3.7). Найкращою моделю є модель згенерована методом AdaBoost за показником $MCC = 0{,}43$. Вона перевершує найкращу з наразі наявних моделей для прогнозування нічної гіпоглікемії DIAppvisor_risk, показник якої $MCC = 0{,}29$, на 48% по показнику MCC.

Проведено аналіз впливу факторів на результати прогнозування. Виявлено, що модель яка вважає показники зросту, ваги, віку і BMI найважливішими, працює краще, ніж модель, яка вважає показники замірів більш важливими. З цього можна зробити висновок, що зріст, вага і вік хворого мають значний вплив на результат прогнозу. Стать, тривалість захворювання і схема лікування хворого майже не мають впливу на результати прогнозування.

ВИСНОВКИ

У дисертаційній роботі одержано такі нові теоретичні та практичні результати:

а) проаналізовано найпопулярніші з наявних методів конструювання дерев прийняття рішень і методи покращення результатів прогнозування дерев прийняття рішень. Обрано методи для подальшої реалізації і обґрунтовано їх вибір;

в) методи було пристосовано до використання на 17 наборах даних з різними включеними стовпцями, які були отримані з оброблених даних DirecNet. До даних DirecNet перед виділенням 17 наборів даних застосовувалася методика виділення ключових значень з показів CGM для прогнозування нічної гіпоглікемії у хворих на цукровий діабет 1-го типу;

г) методи було реалізовано засобами мови програмування R і протестовано на 17 підготованих наборах даних. В результаті роботи моделей можна зробити висновок, що найкраще по показнику MCC серед методів прогнозування на основі дерев прийняття рішень проявив себе метод, заснований на AdaBoost ($MCC = 0{,}43$), а найгірше — на C4.5 ($MCC = 0{,}3$);

д) розроблені методи проявили себе при тестуванні значно краще, ніж існуючі методи. Модель, отримана за допомогою методу AdaBoost ($MCC = 0{,}43$), по показнику MCC на 48% краща за найкращу наявну модель DIAppvisor_risk ($MCC = 0{,}29$);

е) проведено аналіз впливу факторів на результати прогнозування. Виявлено, що модель, яка вважає показники зросту, ваги, віку і BMI найважливішими, працює краще, ніж модель, яка вважає показники замірів більш важливими. Зріст, вага і вік хворого мають значний вплив на результат прогнозу. Стать, тривалість захворювання і схема лікування хворого майже не мають впливу на результати прогнозування.

У подальшому результати роботи методів може бути покращено за рахунок використання даних новіших клінічних досліджень і, бажано, більшого обсягу для навчання моделей.

ПЕРЕЛІК ПОСИЛАНЬ

1. Tan P. Introduction to data mining / P.N. Tan, M. Steinbach, V. Kumar. — 1st ed. — Boston: Pearson, 2005. — 769 p.

2. Rokach L. Data mining and knowledge discovery handbook / L. Rokach, M. Oded. — 2nd ed. — New York: Springer, 2010. — 1285 p.

3. Hancock T. Lower Bounds on Learning Decision. Lists and Trees / T.R. Hancock, T. Jiang, M. Li, J. Tromp // Information and Computation. — 1996. — Vol. 126 (2). — P. 114-122.

4. Hyafil L. Constructing optimal binary decision trees is NP-complete / L. Hyafil, R.L. Rivest // Information Processing Letters. — 1976. — Vol. 5 (1). — P. 15-17.

5. Quinlan J. Induction of decision trees // Machine Learning. — 1986. — Vol. 1 (1). — P. 81-106.

6. Quinlan J. C4.5: Programs for Machine Learning // Machine Learning. — San Mateo: Morgan Kaufmann Publishers, Inc. — 1993. — Vol. 16 (3). — P. 235-240.

7. Breiman L. Classification and Regression Trees / L. Breiman, J. Friedman, R. Olshen, C. Stone // Information and Computation. — Wadsworth Int. Group, 1984. — 368 p.

8. J. Sonquist Searching for Structure / J.A. Sonquist, E.L. Baker, J.N. Morgan. — Ann Arbor: Institute for Social Research, University of Michigan, 1971. — 287 p.

9. Gillo M. MAID: A Honeywell 600 program for an automatised survey analysis // Behavioral Science. — 1972. — Vol. 17. — P. 251-252.

10. Morgan J. THAID: a sequential search program for the analysis of nominal scale dependent variables / J.N. Morgan, R.C. Messenger. — Ann Arbor: Institute for Social Research, University of Michigan, 1973. — Vol. 17. — P. 251-252.

11. Kass G. An exploratory technique for investigating large quantities of categorical data // Applied Statistics. — 1980. — Vol. 29 (2). — P. 119-127.

12. Breiman L. Bagging predictors // Machine learning. — 1996. — Vol. 24 (2). — P. 123-140.

13. Louppe G. Understanding random forests: From theory to practice. — 2014.— 225 p.

14. The Random Forest Algorithm [Електронний ресурс]. — Swiss Federal Institue of Technology Zurich, 2014. — Режим доступу: https://stat.ethz.ch/education/semesters/ss2012/ams/slides/v10.2.pdf

15. Appel R. Quickly boosting decision trees-pruning underachieving features early / R. Appel, T. Fuchs, P. Dollar, P. Perona // JMLR Workshop and Conference Proceedings. — 2013. — Vol. 28. — P. 594-602.

16. Autermann C. Boosted Decision Trees: A modern method of data analysis [Електронний ресурс]. — University Of Hamburg, 2007. — Режим доступу: http://wwwiexp.desy.de/users/auterman/talks/20070706_hh_bdt.pdf

17. Hastie T. Trees, Bagging, Random Forests and Boosting [Електронний ресурс]. — Stanford University, 2003. — Режим доступу: http://jessica2.msri.org/attachments/10778/10778-boost.pdf

18. CRAN Task View: Machine Learning & Statistical Learning [Електронний ресурс]. — Режим доступу: https://cran.r-project.org/web/views/MachineLearning.html

19. Decision Trees — scikit-learn documentation [Електронний ресурс]. — Режим доступу: http://scikit-learn.org/stable/modules/tree.html

20. Frequently Asked Questions on R [Електронний ресурс]. — Режим доступу: https://cran.r-project.org/doc/FAQ/R-FAQ.html

21. Olivetti E. Statistical independence for the evaluation of classifier-based diagnosis / E. Olivetti, S. Greiner Avesani, P. Avesani // Brain Informatics. — 2015. — Vol. 2 (1). — P. 13-19.

22. Tkachenko P. Prediction of Nocturnal Hypoglycemia by an aggregation of previously known prediction approaches: Proof of concept for clinical application / P. Tkachenko, G. Kriukova, M. Aleksandrova, O. Chertov, E. Renard, S. Pereverzyev [Електронний ресурс]. — Режим доступу: http://www.ricam.oeaw.ac.at/files/reports/16/rep16-06.pdf