Міністерство освіти і науки України
Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук
(повна назва)

Кафедра Штучного інтелекту
(повна назва)

# АТЕСТАЦІЙНА РОБОТА
# Пояснювальна записка

Рівень вищої освіти другий (магістерський)
(рівень вищої освіти)

Дослідження та розробка методів прогнозування з використанням імовірнісних нейронних мереж
(тема)

Виконав:
студент 2 курсу, групи СШІм-18-1
Северина С.С.
(прізвище, ініціали)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва спеціальності)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва освітньої програми)

Керівник проф. Бодянський Є.В.
(посада, прізвище, ініціали)

Допускається до захисту

Зав. кафедри _____ В.О. Філатов
(підпис) (прізвище, ініціали)

2019 р.

Харківський національний університет радіоелектроніки

Факультет Комп'ютерних наук

Кафедра Штучного інтелекту

Рівень вищої освіти другий (магістерський)

Спеціальність 122 – Комп'ютерні науки
(код і повна назва)

Тип програми освітньо-професійна
(освітньо-професійна або освітньо-наукова)

Освітня програма Системи штучного інтелекту (СШІ)
(повна назва)

ЗАТВЕРДЖУЮ:
Зав. кафедри _____
(підпис)
«____»_____ 2019 р.

# ЗАВДАННЯ

НА АТЕСТАЦІЙНУ РОБОТУ

студентові Севериній Світлані Сергіївні
(прізвище, ім'я, по батькові)

1. Тема роботи Дослідження та розробка методів прогнозування з використанням імовірнісних нейронних мереж

затверджена наказом по університету від 04 листопада 2019 р. № 1623Ст

2. Термін подання студентом роботи до екзаменаційної комісії 18 грудня 2019 р.

3. Вихідні дані до роботи Науково-технічні публікації, дані Інтернет-джерел та відомих наукових проектів щодо розробки та дослідження імовірнісних нейронних мереж для вирішення задачі прогнозування, тестові вибірки результатів футбольних змагань, деталі технічної реалізації

4. Перелік питань, що потрібно опрацювати в роботі Аналіз методів прогнозування, аналіз футбольної предметної галузі, аналіз імовірнісних нейронних мереж, розробка модуля прогнозування, тестування та аналіз результатів.

5. Перелік графічного матеріалу із зазначенням креслеників, схем, плакатів, комп'ютерних ілюстрацій (п.5 включається до завдання за рішенням випускової кафедри) Рисунок 1 – Методи прогнозування, Рисунок 2 – Моделі прогнозування, Рисунок 3 – Моделі часових рядів, Рисунок 4 – Загальна інформація про матч, Рисунок 5 – Структурна схема нейрона, Рисунок 6 – Приклад букмекерського сайту, Рисунок 7 – Структурна схема ІНС, Рисунок 8 – Інформація по туру, Рисунок 9 – Інформація по грі, Рисунок 10 – Результати пошуку матчів, Рисунок 11 – Точності першого метода для різних типів матчів, Рисунок 12 – Порівняння першого метода з справжніми даними, Рисунок 13 – Точності другого метода для різних типів матчів,

6. Консультанти розділів роботи (п.6 включається до завдання за наявності консультантів згідно з наказом, зазначеним у п.1 )

| Найменування розділу | Консультант (посада, прізвище, ім'я, по батькові) | Позначка консультанта про виконання розділу | |
|---|---|---|---|
| | | підпис | дата |
| Основна частина | проф. Бодянський Є.В. | | |
| | | | |

**КАЛЕНДАРНИЙ ПЛАН**

| № | Назва етапів роботи | Терміни виконання етапів роботи | Примітка |
|---|---|---|---|
| 1 | Отримання завдання на атестаційну роботу | 04.11.19 | виконано |
| 2 | Аналіз предметної області і постановка завдання | 04.11.19-12.11.19 | виконано |
| 3 | Дослідження методів навчання | 13.11.19-14.11.19 | виконано |
| 4 | Створення імітаційної моделі | 15.11.19-19.11.19 | виконано |
| 5 | Тестування і відладка імітаційної моделі | 20.11.19-28.11.19 | виконано |
| 6 | Обробка і оформлення результатів | 29.11.19-03.11.19 | виконано |
| 7 | Оформлення пояснювальної записки | 04.12.19-12.12.19 | виконано |
| 8 | Нормоконтроль | 13.12.19 | виконано |
| 9 | Попередній захист | 16.12.19 | виконано |
| 10 | Захист перед ЕК | 18.12.19 | |

Дата видачі завдання 04 листопада 2019 р.

Студент _____
(підпис)

Керівник роботи _____ проф. Бодянський Є.В.
(підпис) (посада, прізвище, ініціали)

# РЕФЕРАТ

Записка пояснювальна: 73 с., 13 рис., 15 табл., 2 дод.,17 джерел.

БАЗА ДАНИХ, ІМОВІРНІСТЬ, МЕТОД, МОДЕЛЬ, НЕЙРОННА МЕРЕЖА, ПРЕПРОЦЕСІНГ, ПРОГНОЗУВАННЯ, ФУТБОЛ

Об'єкт дослідження – методи соціально-економічного прогнозування.

Предмет дослідження – статистична інформація результатів футбольних матчів.

Метою даної атестаційної роботи є розробка методів прогнозування результатів футбольних матчів з використанням імовірнісних нейронних мереж, оцінки їх точності і отримання висновку про можливість його застосування до футбольної галузі.

Методи дослідження – статистичні, системного аналізу, оптимізаційні.

# РЕФЕРАТ

Пояснительная записка: 73 с., 13 рис., 15 табл., 2 прил., 17 источников.

БАЗА ДАННЫХ, ВЕРОЯТНОСТЬ, МЕТОД, МОДЕЛЬ, НЕЙРОННАЯ СЕТЬ, ПРЕПРОЦЕССИНГ, ПРОГНОЗИРОВАНИЕ, ФУТБОЛ

Объект исследования – методы социально-экономического прогнозирования.

Предмет исследования - статистическая информация результатов футбольных матчей.

Целью данной аттестационной работы является разработка методов прогнозирования результатов футбольных матчей с использованием с использование вероятностной нейронной сети, оценка точности и получение вывода о возможности их применения к футбольной отрасли.

Методы исследования – статистические, системного анализа, оптимизационные.

## ABSTRACT

Explanatory note: 73 pages, 13 figures, 15 tables, 17 sources, 2 annex.

DATABASE, FOOTBALL, FORECASTING, METHOD, MODEL, NEURAL NETWORK, PREPROCESSING, PROBABILITY

Object of research: methods of social-economic forecasting.

Subject of research - statistical information on the results of football matches.

The purpose of this qualification work is to develop methods for predicting the results of football matches using a probabilistic neural network, assess accuracy and evaluate the feasibility of their application in football area.

Research methods: statistics, system analysis, optimization.

**ЗМІСТ**

# ПЕРЕЛІК УМОВНИХ ПОЗНАЧЕНЬ, СИМВОЛІВ, ОДИНИЦЬ, СКОРОЧЕНЬ І ТЕРМІНІВ

БД – база даних;

Дестимулятори – фактор, при збільшенні якого ймовірність перемоги зменшується;

ІНМ – імовірнісна нейронна мережа;

Сейв – ситуація під час матчу, коли дії воротаря запобігають голу;

Стимулятор – це фактор, при збільшенні якого ймовірність перемоги збільшується;

ШНМ – штучна нейронна мережа.

# ВСТУП

Події, що відбуваються навколо нас, залежать від безлічі факторів, деякі з яких є випадковими, а інші – ні. Людський мозок здатний вловлювати взаємозв'язок між невеликою кількістю умов. При наявності багатьох критеріїв вибору людині стає складно або навіть неможливо зробити оптимальний вибір, тобто коли мова йде про прогнозування результатів складних подій, наприклад, ціни на біржі, курс валют, політична ситуація в світі. Майбутнє залежить від безлічі інших подій, які певним чином співвідносяться один до одного. Деякі події мають більш вагомий вплив, деякі – слабший. Пошуком ваги впливу події на результат займаються експерти предметної галузі. Проте в більшості випадків прогнозування життєвих подій ґрунтується на великій кількості умов, а їх вплив не є постійним. Для збереження актуальності прогнозу, необхідна велика кількість людських ресурсів. У таких випадках на допомогу людині приходить комп'ютер, з його здібностями за лічені секунди проводити тисячі математичних і логічних операцій, не загубивши ніяких деталей.

Для прогнозування використовують системи з великою кількістю вагомих вхідних даних, нерідко використовують моделі предметної галузі. Моделі предметної галузі – це такі математичні моделі прогнозування, для побудови яких використовують закони предметної області. Зазвичай для побудови такої моделі необхідна участь експерта. Проте існують задачі, наприклад, прогнозування результатів футбольних матчів, для яких складно, або навіть неможливо, побудувати універсальну модель.

Щодо футбольних матчів, це пов’язано з двома факторами. По-перше, модель буде залежати щонайменше від команди, а в складних випадках від її гравців, по-друге, динаміка розвитку буде потребувати регулярного переналаштування коефіцієнтів. Наприклад, існує низка кількісних параметрів для опису результатів гри, серед яких відсоток володіння м’ячем [1], котрий для одного матчу має прямопропорційний вплив на

результату, а для іншого – зворотньопропорційний. Як наслідок маємо необхідність у послідовному online режимі змінювати коефіцієнти впливу кожного параметра на результат для команди. Для автоматизації цієї задачі пропонується використовувати імовірнісну нейронну мережу.

Дана робота спрямована на дослідження прогнозування результатів футбольних матчів за допомогою імовірнісної нейронної мережі, її програмна реалізація засобами Matlab. Остання частина присвячена порівнянню якості прогнозу, залежно від об'єму навчальної вибірки та різних підходів до препроцесінгу вхідного образу.

# 1 АНАЛІЗ ПРЕДМЕТНОЇ ГАЛУЗІ ТА ФОРМАЛІЗОВАНА ПОСТАНОВКА ЗАДАЧІ

Кожен день десь в світі проходять спортивні змагання. Футбол, баскетбол, хокей, волейбол – це тільки мала частина спортивного світу. Мільйони людей щодня вболівають за свою улюблену команду або спортсмена, і будь-який уболівальник хоче, щоб його сторона перемогла. За даними досліджень, в яких порівнювали 25 видів спорту з 13 критеріями, найпопулярнішим спортом є футбол: більш 3,5 мільярда фанатів по всьому світу [1]. Також футбол є лідером за кількістю ставок на результати матчів.

Всю історію, починаючи з Римської імперії, в ногу зі спортом йде букмекер. Букмекер – це професійний сперечальник, який займається прийомом грошових коштів (ставок) на події (зазвичай спортивні) з попередньо обговореними можливостями, що визначають коефіцієнти, а також розмір виплати в разі перемоги. Очевидно, що букмекер завжди хоче отримувати прибуток, а не збиток, після кожного спортивного змагання. Тому йому необхідно складати прогноз, який би враховував показники успішності гравців змагання. Для змагань, які не мають безліч факторів впливу, можна використовувати інтуїтивні методи, а для складних змагань необхідна побудова моделі прогнозування, яка зможе враховувати всі відомі фактори.

На сьогоднішній день ставки на спорт мають завидну популярність. Запит «ставки на спорт» в пошуковій системі Google видав 28 мільйонів результатів на російській мові і 116 мільйонів англійською. Для порівняння пошукової запит «спортзал» дає 7 і 762 мільйон відповідно. Хоча прихильників занять спортом багато, ця статистика свідчить про те, що користувачі мають потребу в спортивних прогнозах. У цих запитах міститься сотні різних сайтів, які візьмуть у користувача ставки і їх кількість продовжує зростати. Кожен день в рекламі, від друзів або перехожих ви можете почути про появу нової букмекерської організації. Це також

свідчить про те, що букмекерська діяльність має перспективи для зростання. Кожен букмекер має певний спосіб розрахунку коефіцієнтів, який тримає в секреті [2]. Хтось спирається на думку експертів, інші ж вдаються до різних методів прогнозування результатів.

Отже, має сенс розробити власні метод прогнозування, який буде розраховувати ймовірність перемоги/поразки команди, і використовувати його для прогнозу результату і/або обчислення коефіцієнтів при прийомі ставок.

## 1.1 Аналіз предметної області

У даній роботі накладаються дві абсолютно різні предметний області: прогнозування і футбол. Буде доцільним досліджувати кожну з них, щоб мати повне уявлення про всі об'єкти, правила і нюанси.

Ізольованого аналізу даних галузей буде недостатньо для створення модуля прогнозування футбольних матчів. Одним з важливих результатів, які я б хотіла отримати в ході виконання роботи – це порівняння різних підходів прогнозування для футбольних матчів. Можливо, деякі прогнози будуть мати дуже низьку точність, що свідчить про непридатність методу прогнозування. Тому другий розділ роботи буде присвячено ретельному дослідженню алгоритмів прогнозування та їх застосування і теоретичної точності для передбачення результатів ігор.

### 1.1.1 Алгоритми прогнозування

Кожній людині хочеться мати уявлення про те, яким буде його майбутнє, що буде відбуватися навколо. Звичайно ж не представляється можливим заглянути в майбутнє, але припустити, іншими словами спрогнозувати ймовірний вихід, можливість є.

Прогноз – це ймовірнісне судження про майбутній стан об'єкта дослідження. Система прогнозування – це впорядкована сукупність методик, технічних засобів, призначена для прогнозування складних явищ або процесів. На сьогоднішній день є безліч різних методів прогнозування, класифікація яких буде розглянута далі.

Для формалізації класифікації алгоритмів прогнозування варто виділити два основних поняття, такі як: модель і метод прогнозування. Метод прогнозування являє собою послідовність дій, які потрібно зробити для отримання моделі прогнозування. Модель прогнозування є функціональне уявлення, яке адекватно описує досліджуваний процес, і є основою для отримання його майбутніх значень.

На підставі вхідних даних прогнозованого процесу і отриманої моделі отримують вихідні дані, які являють собою судження про певну подію майбутнього. Всі методи прогнозування поділяються за джерелом отримання прогнозу (рисунок 1.1).

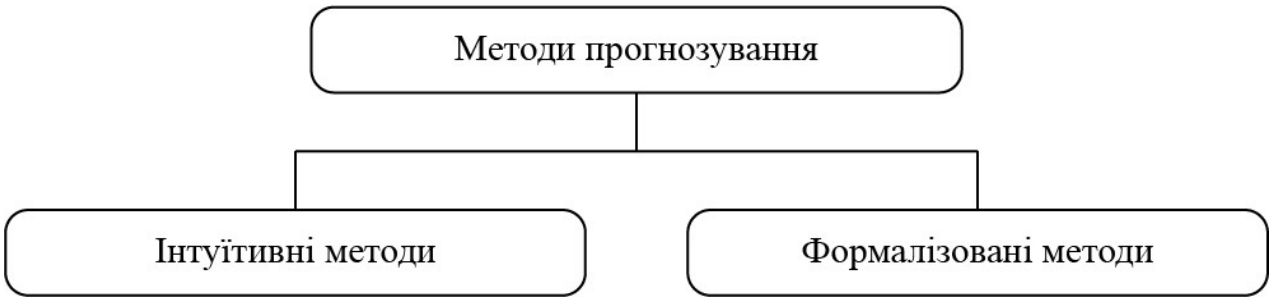


Рисунок 1.1 – Методи прогнозування

Інтуїтивні методи прогнозування мають справу з судженнями та оцінками експертів. Такі методи застосовуються для прогнозування поведінки об'єктів, які не піддаються математичній формалізації або занадто складні для побудови математичної моделі їх функціонування. Найчастіше такі методи не спроможні для створення високоточного прогнозу. Формалізовані методи – описані в літературі методи прогнозування, в

результаті яких будують моделі прогнозування, тобто визначають таку математичну залежність, яка дозволяє визначити поведінку об'єкта – зробити прогноз (рисунок 1.2).


Рисунок 1.2 – Моделі прогнозування

Моделі предметної області – це такі математичні моделі прогнозування, для побудови яких використовують закони предметної області. На підставі висновків предметної області будується модель поведінки об'єкта і моделюється його майбутня поведінка. Моделі часових рядів – математичні моделі прогнозування, які прагнуть знайти залежність поведінки об'єкта від минулого стану всередині самого процесу і на цій залежності обчислити прогноз. Ці моделі можуть прогнозувати поведінку будь-якого об'єкта, стан якого можна описати чисельними характеристиками, при цьому сама модель не змінює свого зовнішнього вигляду, а змінюються лише параметри налаштування моделі прогнозування (рисунок 1.3)[3].

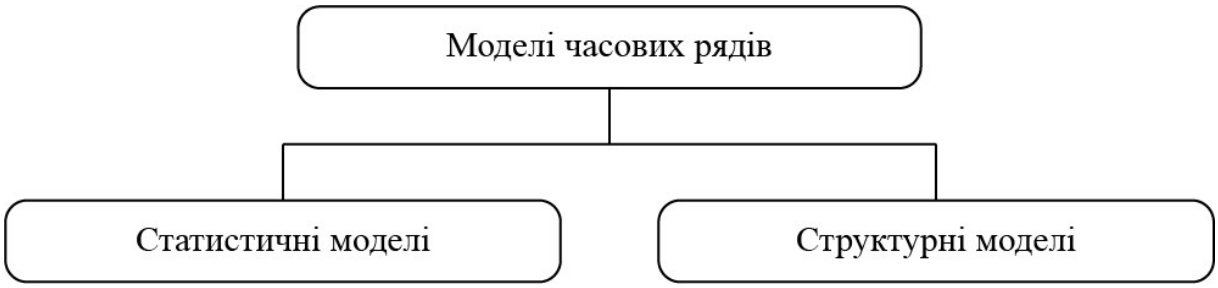


Рисунок 1.3 – Моделі часових рядів

У статистичних моделях залежність майбутнього значення від минулого задається в вигляді деякого рівняння. До них відносяться:

а) регресивні моделі (лінійна регресія, нелінійна регресія);

б) авторегресійні моделі (ARIMAX, GARCH, ARDLM);

в) модель експоненціального згладжування;

г) модель за вибіркою максимального подібності.

У структурних моделях залежність майбутнього значення від минулого задається у вигляді певної структури і правил переходу по ній. До них відносяться:

а) нейромережеві моделі;

б) моделі на базі ланцюгів Маркова;

в) моделі на базі класифікаційно-регресійних дерев.

При прогнозуванні значення можна виділити три основних компоненти: вхідні дані, метод прогнозування і вихід (результат).

Для перетворення входів до виду, необхідному для алгоритму, використовуються різні методи препроцессінга, який створюються на підставі знань експертів предметної області. Залежно від методу прогнозування використовується два типи вихідних даних: перелік результатів матчів і повна результуюча статистика по командам. Прогнозування можливо двох видів: прогноз результату матчу і прогноз підсумкових результатів по турніру. Очевидно, що вихідні дані будуть мати різний тип.

### 1.1.2 Футбольна область

Футбол – командний вид спорту, в якому метою є забити м'яч у ворота суперника ногами або іншими частинами тіла (крім рук) більшу кількість разів, ніж команда суперника. Футбольний матч – це гра між двома футбольними командами за певними правилами. Дані, які несе в собі кожен, матч дуже об'ємні, а їх обробка – трудомісткий процес. Тому автоматизація

роботи з цими даними є необхідним завданням.

Кожен зіграний матч несе в собі безліч різної інформації, а саме:

а) інформація про матчі в цілому (що грають команди, рахунок, дата, стадіон);

б) статистика за гру по кожному гравцеві (забиті голи, отримані картки, кількість часу проведеного в грі);

в) дані про тренерів граючих команд.

В текстовому вигляді інформацію про кожен матч можна представити таким чином (рисунок 1.4).

22 грудня 2017, Арсенал 3 : 3 Ліверпуль
Голи: Коутіньо Філіппе 26’, Салах Мохамед 52’, Санчес Алексіс 53’, Джака Граніт 56’, Озіл Месут 58’, Фірміно Роберто 71’
Стадіон: Емірейтс (Лондон)
Глядачі: 40 000
Суддя: Мартин Аткінсон
Склади команд:
Арсенал: Чех, Бельєрін, Кошельни, Мустафі (Шкодран, 46’), Монреаль, Джака, Вілшер, Санчес (Тео, 89’), Озіл, Івоби (Денни, 78’), Ляказетт.
Ливерпуль: Міньйоле, Гомес, Ловрен, Клаван, Робертсон, Эмре, Хендерсон (Мілнер, 13’), Коутіньо (Окслейд-Чемберлен, 84’), Салах, Фірміно, Мане
ЖК: Івоби.

Рисунок 1.4 – Загальна інформація про матч

Перший рядок несе в собі інформацію про двох командах, між якими проходив матч, з вказаними назвами і містами, при необхідності, і про результат гри. Команду, яка приймала у себе суперницю, прийнято записувати першої. Наступні рядки повідомляють про склад обох команд, причому порядку: воротар, захисники, півзахисники, нападники, і про їхні досягнення в грі (голи картки, кількість проведених на полі хвилин).

Однак більш популярною серед футбольних фанатів є інформація не про окремому матчі або команді, а підсумкова таблиця результатів певного чемпіонату, яку, зазвичай представляють у вигляді рейтингового списку досягнень команд (таблиця 1.1)[9].

Таблиця 1.1 – Підсумки сезону

| М | Команда | І | В | Н | П | М’ячі | О |
|---|---|---|---|---|---|---|---|
| 1 | Манчестер Сіті | 38 | 32 | 4 | 2 | 106:27 | 100 |
| 2 | Манчестер Юнайтед | 38 | 25 | 6 | 7 | 68:28 | 81 |
| 3 | Тоттенхем | 38 | 23 | 8 | 7 | 74:36 | 77 |
| 4 | Ліверпуль | 38 | 21 | 12 | 5 | 84:38 | 75 |
| 5 | Челсі | 38 | 21 | 7 | 10 | 62:38 | 70 |
| 6 | Арсенал | 38 | 19 | 6 | 13 | 74:51 | 63 |
| 7 | Бернлі | 38 | 14 | 12 | 12 | 36:39 | 54 |
| 8 | Евертон | 38 | 13 | 10 | 15 | 44:58 | 49 |
| 9 | Лестер | 38 | 12 | 11 | 15 | 56:60 | 47 |
| 10 | Ньюкасл | 38 | 12 | 8 | 18 | 39:47 | 44 |
| 11 | Крістал Пелас | 38 | 11 | 11 | 16 | 45:55 | 44 |
| 12 | Борнмут | 38 | 11 | 11 | 16 | 45:61 | 44 |
| 13 | Вест Хем | 38 | 10 | 12 | 16 | 48:68 | 42 |
| 14 | Уотфорд | 38 | 11 | 8 | 19 | 44:64 | 41 |
| 15 | Брайтон | 38 | 9 | 13 | 16 | 34:54 | 40 |
| 16 | Хаддерсфілд | 38 | 9 | 10 | 19 | 28:58 | 37 |
| 17 | Саутгемптон | 38 | 7 | 15 | 16 | 37:56 | 36 |
| 18 | Суонсі | 38 | 8 | 9 | 21 | 28:56 | 33 |
| 19 | Сток Сіті | 38 | 7 | 12 | 19 | 35:68 | 33 |
| 20 | Вест Бромвіч | 38 | 6 | 13 | 19 | 31:56 | 31 |

Таблиця відображає кількість зіграних матчів (М), яке на кінець туру повинно бути однаково для всіх команд, якщо не відбулося якихось форс-мажорних обставин. Також матчі, які були виграні (В), зіграні в нічию (Н) і програні (П). Окрема колонка виділена під співвідношення між забитими і пропущеними голами.

Записи в таблиці упорядковано відповідно до зменшенням кількості набраних командою очок, які розраховуються за формулою 1.1:

$$О = В \cdot 3 + Н \cdot 1 + П \cdot 0, \quad (1.1)$$

де О – сумарна кількість очок,

В – кількість виграних матчів,

Н – кількість матчів зіграних у нічию,

П – кількість поразок.

Однак інформація про матчі в такому вигляді незручна і складна в обробці, особливо якщо її необхідно відображати і обробляти більш ретельно. Тому доцільніше зберігати інформацію в табличному вигляді, представленому на рисунку 1.4, так як тільки в такому вигляді можна гарантувати валідність, несуперечливість та цілісність даних. Також зберігання даних в такому вигляді спростить обробку: однією функцією запиту до бази даних можна отримати таблицю 1.1.

## 1.2 Нейронні мережі

Штучну нейронну мережу (ШНМ) формально можна визначити, як пару (N, E), де N – множина нейронів, а E – множина зв'язків між ними. Нейрон з біологічної точки зору є біологічною системою, призначеної для передачі і обробки інформації. Штучний нейрон – це елемент ШНМ, який являє собою аналог біологічного нейрона. У загальному випадку на вхід нейрона надходить вектор сигналів X, кожен елемент якого має власний

ваговий коефіцієнт. Далі, підсумовуючи отримані в ході перетворень значення, вони надходять в функцію активації, яка перетворює її в вихідний сигнал. Структурна узагальнена схема найпростішого нейрона представлена на рисунку 1.5 [5].

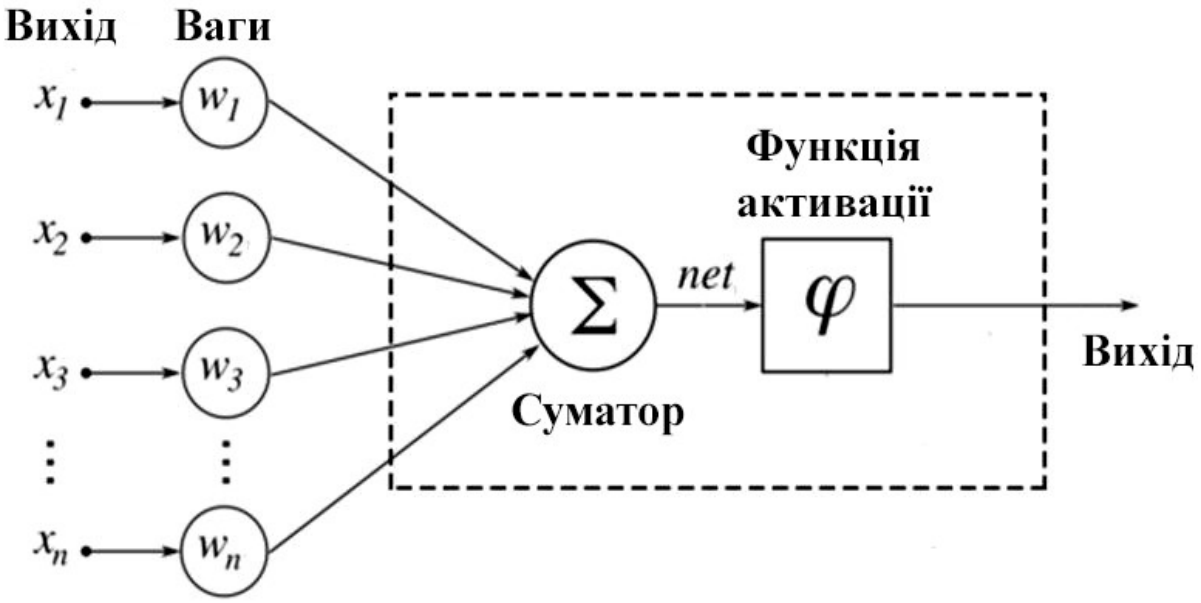


Рисунок 1.5 – Структурна схема нейрона

Пов'язуючи між собою множину нейронів, утворюється нейронна мережа. При побудові ШНМ на кожному шарі нейронів може знаходиться довільна кількість нейронів з різними активаційними функціями. Вибір кількості шарів, типів нейронів і зв'язків залежить від розв'язуваної задачі, наприклад, прогнозування, класифікація або кластеризація.

При вирішенні задачі класифікації можна оцінити щільність ймовірності для кожного класу, порівняти між собою ймовірності приналежності до різних класів і вибрати найбільш ймовірний. Насправді саме це відбувається, коли ми навчаємо нейронну мережу вирішувати задачу класифікації – мережа намагається визначити (тобто апроксимувати) щільність ймовірності [4].

Імовірнісна нейронна мережа (ІНМ) – це нейронна мережа з прямим зв'язком, яка широко використовується в задачах класифікації та розпізнавання образів. ІНМ відносяться до ШНМ з прямою передачею інформації і її архітектура подібна до радіально-базисних та узагальнених регресійних мереж [5].

В алгоритмі ІНМ функція розподілу імовірності (ФРІ) кожного класу апроксимується оцінкою Парзена, що використовує вагові функції, які мають центр в точках відповідних образів з відомою класифікацією з навчальної вибірки та непараметричної функцією. Потім, використовуючи ФРІ кожного класу, оцінюється імовірність приналежності до класу для нових вхідних даних. Після, використовуючи правило Байєса, виділяють клас з найбільшою апостеріорної імовірністю для поточних вхідних даних. Завдяки цьому методу імовірність помилкової класифікації зводиться до мінімуму. Цей тип ШНС було отримано з байєсівської мережі та статистичного алгоритму, званого дискримінантним аналізом Кернела Фішера. Він був введений Дональдом Шпехтом в 1966 році.

Дана мережа складається з вхідного шару, першого прихованого – шара образів, другого прихованого – шар підсумовування, і вихідного шару, утвореного в даному випадку одним нейрономкомпоратором.

Ідея байєсівської класифікації полягає в тому, що для кожного вхідного образу можна прийняти рішення на основі вибору найбільш ймовірного класу з тих, яким міг би належати даний образ. Це рішення, однак, вимагає оцінки функції щільності ймовірностей для кожного класу, яка виконується на основі аналізу даних з навчальної вибірки. Дана обставина обмежує процес навчання імовірнісних мереж тільки пакетним режимом.

ІНМ призначені для вирішення завдань байєсівської класифікації (розпізнавання типу образа, враховуючи рівень подібності на типізовані образи), в основі якої лежить одна з основних теорем теорії ймовірностей, а саме формула Байеса (формула 1.2):

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}, \quad (1.2)$$

де P(A) – апріорна ймовірність події A;

P(B) – апріорна ймовірність події B;

P(A|B) – ймовірність настання події A, при умові настання події B;

P(B|A) – ймовірність настання події B, при умові настання події A.

Сенс цього виразу полягає в тому, що для події A з відомою ймовірністю P(A), умовна ймовірність P(A|B) може бути обчислена на основі так званої апостеріорної ймовірності P (B|A) і ймовірностей подій P(A) і P(B). З позиції завдання класифікації тут Y інтерпретується як можливий клас, в який може потрапити класифікується образ, а X розглядається як власне вхідний вектор-образ.

## 1.3 Порівняльна характеристика конкурентної продукції

Як уже неодноразово було сказано, в сучасному світі ставки на спорт є популярною розважальною галуззю для громадян багатьох країн. Так як потенційно створювана програма має бути випущена для загального користування, то вона повинна бути конкурентоспроможним. Даний пункт якраз і присвячений аналізу наявних аналогічних сервісів. Приклад сайту для ставок можна побачити на рисунку 1.6.

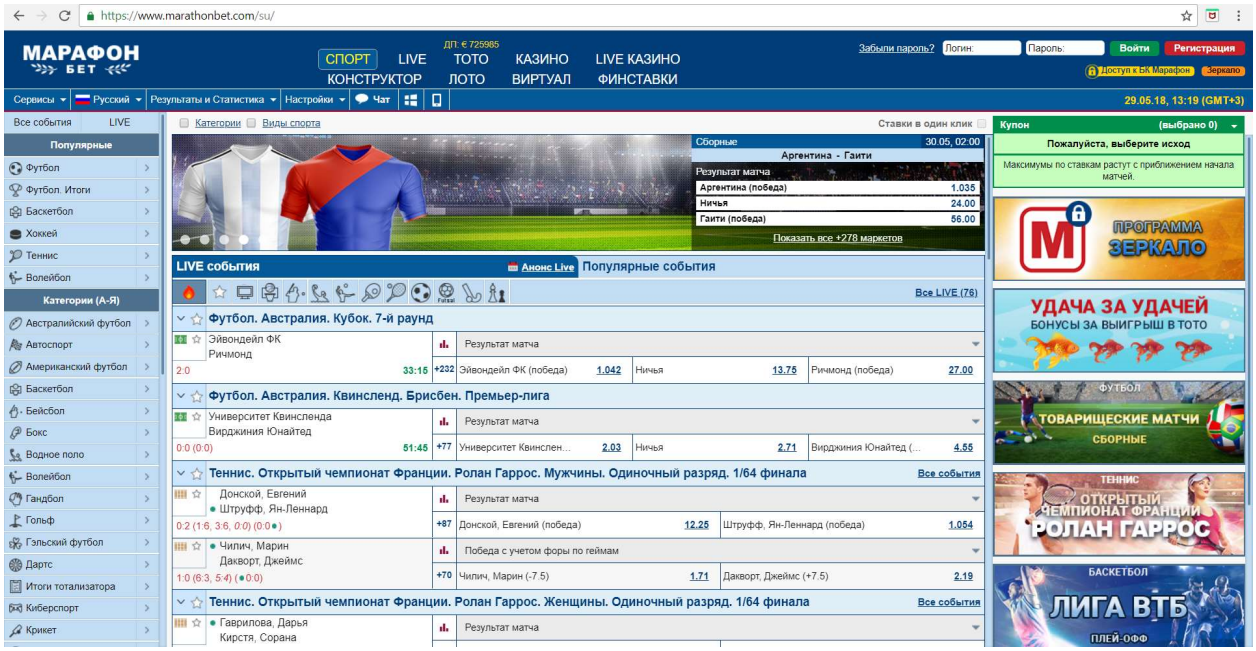


Рисунок 1.6 – Приклад букмекерського сайту

Для модулів, які прогнозують футбольні матчі, існує два протилежних способу монетизації: через продаж прогнозів і через отримання ставок (обчислення коефіцієнтів для результатів). Очевидно, що дві ці області тісно взаємодіють один з одним. Однак, обидва види сервісів приховують те, яким чином було отримано їх прогноз або коефіцієнт [7]. Тому порівнювати існуючі сайти за точністю їх прогнозів не представляється можливим.

Так як на даний момент існують сотні різних сайтів для ставок на спорт, то, крім порівняння з основного функціоналу, має місце бути порівняння з «зручності» найпопулярніших з них в нашому регіоні. Порівняння за основними характеристиками представлено у таблиці 1.2.

Таблиця 1.2 – Порівняльна характеристика букмекерських сайтів

| Букмекер | Мови | Бонус | Мобільна версія | Платіжна система | | | | | | | Служба підтримки | | | Оцінка користувачів |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | | VISA | WebMoney | Яндекс.Гроші | Термінал | Приват24 | Bitcoin | Qiwi | Живий чат | Електронна пошта | Телефон | |
| 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 | 9 | 10 | 11 | 12 | 13 | 14 | 15 |
| Марафон | 26 | + | + | + | + | + | + | – | – | – | + | + | + | 5 |
| Пари-Матч | 6 | + | + | + | + | + | – | – | – | + | – | + | + | 4,87 |
| GameBet.com | 2 | + | + | + | + | + | – | + | + | + | + | – | + | 4,36 |
| 1xbet | 40 | + | + | + | + | + | – | – | – | + | + | + | + | 4,76 |
| Titanbet | 5 | + | + | + | + | – | – | – | – | + | + | + | + | 4,12 |
| Leon | 5 | + | + | + | + | + | – | – | – | + | – | + | – | 4,03 |
| Фаворит | 2 | – | + | + | – | – | + | + | – | – | + | + | + | 3,1 |
| Winline | 1 | + | + | + | – | + | – | – | – | + | + | + | + | 4 |
| Вулкан | 2 | + | + | + | + | + | + | – | + | + | + | + | – | 4,26 |
| Лига Ставок | 2 | + | + | + | + | + | + | – | – | + | + | + | + | 3,79 |
| SportingBet | 23 | + | + | + | + | + | – | – | – | – | + | + | + | 3,33 |

У таблиці 1.2 наведено порівняння невеликої кількості онлайн-букмекерів за основними характеристиками, які важливі для користувача. Так як ставки в інтернеті популярні вже не перший рік, то очевидно, що розвиток сайтів для них на гідному рівні.

Для користувачів є безліч зручних способів оплати, наявність мобільної версії, що в даний час є вагомим фактором, так як люди все частіше використовують мобільні телефони замість персональних комп'ютерів і ноутбуків. Також для користувача важливо, як швидко технічна підтримка зможе вирішити виниклі труднощі при роботі з сервісом. Важливим фактором успіху будь-якого проекту є обсяг потенційного ринку користувачів і кількість мов чудово характеризує сервіси за цією ознакою.

Для того, щоб мати можливість скласти конкуренцію на даному ринку, необхідно докласти достатню кількість зусиль при розробці сайту: користувач повинен захотіти використовувати саме цей сервіс. Для досягнення такого ефекту необхідно, щоб кінцевий продукт не поступався, а краще перевершував існуючі альтернативи.

## 1.4 Постановка задачі

Метою даної атестаційної роботи є розробка алгоритму прогнозування результатів футбольних матчів з використанням ймовірнісної нейронної мережі.

Для реалізації мети на основі проведеного аналізу предметної області можна чітко сформулювати чотири етапів виконання атестаційної роботи магістра:

а) провести детальний аналіз футбольної області, побудова моделі даних;

б) провести всебічний аналіз існуючих алгоритмів прогнозування;

в) розробити і порівняти алгоритми прогнозування результатів футбольних матчів та підсумкової турнірної таблиці;

г) підготувати матеріали для захисту атестаційної роботи.

Планування і тайм-менеджмент – невід'ємна частина будь-якого успішного проекту, будь то корпоративна завдання, навчальне завдання або просто справи по дому. Тому для досягнення найкращого результату виконання атестаційної роботи, кожний з перерахованих етапів доцільно конкретизувати для повного розуміння майбутньої роботи.

Перший етап – детальний аналіз футбольної галузі. Тут можна визначити наступні пункти:

а) виділити основні сутності;

б) визначити залежності між сутностями;

в) побудувати цілісну базу даних.

Другий етап – аналіз існуючих алгоритмів прогнозування. Попереднє дослідження алгоритмів було описано в даному розділі, але його, очевидно, недостатньо для програмної реалізації. На даному етапі можна визначити наступні пункти:

а) розробити математичне представлення задачі прогнозування футбольних матчів;

б) дослідити методи прогнозування, що ґрунтуються на моделі предметної області;

в) дослідити методи прогнозування часових рядів;

г) дослідити нейронні мережі в цілому та ретельно імовірнісну нейронну мережу;

д) сформулювати математичну постановку задачі для прогнозування результатів футбольних матчів за допомогою ІНМ;

е) теоретично оцінити точність розглянутих алгоритмів.

Третій етап – це розробка алгоритмів прогнозування футбольних матчів. В цей етап можна включити наступні пункти:

а) виконати препроцессінг футбольної статистики;

б) реалізувати алгоритми прогнозування результатів футбольних матчів за допомогою ІНМ;

в) провести тестування алгоритмів, аналіз і порівняння отриманих результатів.

П'ятий і заключний етап виконання роботи містить в собі кілька пунктів, по закінченню виконання яких атестаційна робота буде завершена:

а) підготувати пояснювальну записку;

б) пройти перевірки оформлення та плагіату;

в) презентацію та доповідь для захисту атестаційної роботи.

# 2 ТЕОРЕТИЧНІ ДОСЛІДЖЕННЯ

У першому розділі був проведений загальний аналіз двох предметних областей, які перетинаються в даній роботі, і порівняльний аналіз аналогів на існуючому ринку. Другий розділ пояснювальної записки буде присвячений теоретичним дослідженням всіх об'єктів, що стосуються даної роботи, і їх властивості.

Перш ніж приступати до розробки алгоритмів прогнозування матчів, необхідно виділити основні сутності футбольної області і правила взаємодії між ними. Вихідні дані алгоритму прогнозування – це результат матчу. Після розробки схеми БД з'явиться чітке уявлення про вхідних даних алгоритмів і можна буде сформулювати завдання прогнозування футбольних матчів, а потім описати застосування методів прогнозування різних видів, а саме: з використанням часових рядів і побудовою моделі предметної області [3]. Опис дасть можливість теоретично оцінити придатність виборних методів до футбольної предметної області. Даний розділ відповідає першому і другому етапу виконання атестаційної роботи.

## 2.1 Футбольна теорія

Футбол бере свої витоки ще в 19 столітті і перші офіційні футбольні правила були опубліковані в 1863 році. Останній несуттєві поправки в сучасну версію футбольних правил були внесені в 1990 році [8]. В даний час правила містить в собі 17 пунктів, які регламентують футбольний матч: поле, м'яч, тривалість гри, порушення, склади команд і т.п. На підставі цих правил і деяких невказаних в них особливостей футболу будуть виділені основні реальні об'єкти (сутності), проведено їх абстрагування, шляхом виділення і фіксації набору їх властивостей, визначення поведінки об'єктів і взаємозв'язків між ними.

Головними сутностями обраної предметної області є:

а) матч;

б) команда;

в) футболіст;

г) позиція;

д) подія в грі;

е) тренер;

ж) турнір;

з) стадіон.

Для кожної сутності є якісь особливості і правила, які в майбутньому будуть застосовані як обмеження цілісності при створенні бази даних, і будуть розглянуті нижче.

Сутність «Матч» необхідна для зберігання загальної інформації про гру. У кожній грі має бути дві різні команди, за винятком ситуації, коли змагаються різні склади однієї команди, але в сучасних чемпіонатах таке зустрічається мало і частіше це відбувається в форматі неофіційних зустрічей. Кожна зареєстрована гра відбувається в рамках будь-якого турніру (Англійська Прем'єр-ліга, Ла Ліга, Ліга Чемпіонів, Бундесліга і т.д.) або є товариським матчем. Матч проходить на певному стадіоні і в залежності від чемпіонату вимоги до розмірів стадіону посилюються (Правило 1) [8]. Дата і час гри визначається заздалегідь і може бути змінена в разі непередбачених обставин.

Кількість гравців, які заявлені на гру, встановлюється регламентом змагань і може варіюватися від 14 до 23. Одночасно на полі знаходяться не більше 11 гравців, а в матчі з урахуванням замін можуть приймати від 11 гравців до 17 гравців, в залежності від допустимої кількості замін. Зворотні заміни допускаються тільки на низькорівневих чемпіонатах. Кількість гравців по позиціях не регламентується, за винятком тільки одного воротаря, який протягом гри може змінюватися з будь-яким гравцем на полі

і це не є заміною. Гравець буде видалений з поля без права заміни (команда грає в меншості), якщо він отримав червону картку [8].

Після закінчення матчу сутність доповнюється статистичною інформацією по матчу, яка зацікавить футбольних фанатів і знадобиться для прогнозування результатів майбутніх ігор:

а) рахунок (кількість забитих голів обома командами);

б) отримані картки;

в) кількість порушень;

г) кількість ударів по воротах;

д) кількість ударів в створ воріт;

е) кутові і штрафні удари;

ж) відсоток володіння м'ячем (сума по двом командам повинна дорівнювати 100%);

з) сейви.

Сутність «Команда» зберігає основну інформацію про набір гравців, об'єднаних разом в футбольний клуб. Кожна команда має назву, місто і сайт. Назва і місто можуть бути замінені на країну, якщо мова йде про національні збірні. Команда може приймати участь в різних турнірах, якщо це дозволяє регламент турніру, наприклад, в національних турнірах можуть брати участь тільки клуби – цієї країни. Кілька сутностей «Тренер» може бути пов'язані з командою в певний момент часу. Також і з гравцями: в команді можуть значитися десятки гравців, але в матчі брати участь не більше 23.

Сутність «Футболіст» описує загальні дані гравців: прізвище, ім'я, дата народження. Її можна розширити іншими біографічними даними, але для даної роботи таких даних достатньо. Останній атрибут сутності – позиція. Він може мати тільки чотири значення: воротар, захисник, півзахисник і нападаючий. Але під час матчу гравець може знаходитися в іншій позиції, яка залежить від обраної тактичної схеми побудови гравців, наприклад, зараз популярною є схема 4-4-2, де присутні 4 захисника, 4 півзахисники і 2 нападники.

Сутність «Позиція» є допоміжною для футболіста. Вона формально вказує звичайне положення гравця на поле.

Сутність «Події в грі» встановлює перелік можливих дій футболістів, які повинні бути зафіксовані. По суті дана сутність є словником, за допомогою якого зв'язуються сутності «Футболіст» і «Матч». Деякі події можу повторюватися, наприклад, реалізовані або нереалізовані голи, порушення, інші ж навпаки відбуватися тільки один раз, наприклад, вийшов в основному складі, вийшов на заміну, вилучений з поля.

Сутність «Тренер» за своїми атрибутами схожа на сутність «Футболіст». Містить інформацію про ім'я, прізвище і дату народження. Очевидно, що сутність тренер не має сенсу без команди. Один тренер може тренувати одноразово не більше одне команди.

Сутність «Турнір» – це чемпіонат, в рамках якого відбувається певна кількість матчів. Турніри можуть бути:

а) національними;

б) міжнародними.

Більшість турнірів проводяться з періодичністю раз на рік. Це називається сезоном змагань. В одному турнірі участь приймає обмежена кількість команд. Число команд в турнірі залежить від його регламенту. Турнірна сітка може будуватися за двома принципами: груповий етап (кожна команда грає з кожною) і за олімпійською системою.

Сутність «Стадіон» має назву, адресу, кількість глядачів, розміри поля. Як згадувалося на початку розділу правилами регламентується розмір поля для різного виду турнірів, тому не на будь-якому стадіоні можуть проходити змагання високого рівня.

Після виділення основних сутностей предметної області, на їх основі створимо схему реляційної бази даних. Для кожної сутності створюється своя таблиця в БД, а властивості об'єктів стають атрибутами (стовпчиками). На кожен атрибут накладаються обмеження: унікальність, тип даних, розмір або більш складні обмеження, наприклад, відповідність регулярному

виразу. Так як при програмної реалізації БД краще використовувати англійські назви атрибутів для поліпшення читабельності коду, то і при розробці використовуємо їх.

При ретельному розборі кожної з сутностей неважко помітити наявність зв'язків багато до багатьох. Для вирішення таких ситуацій використовується підхід, при якому створюється допоміжна таблиця. Тому порядок опису розробки БД буде відрізнятися від опису сутностей і починатися з таблиць-словників.

## 2.2 Прогнозування матчів

У першому розділі були розглянуті загальні терміни прогнозування, представлена одна з їх можливих класифікації та наведено кілька варіантів методів. Даний пункт присвячений формулюванні математичної постановки задачі прогнозування футбольних матчів і конкретизації методів для побудови прогнозу.

Футбольна предметна області зберігає в собі безліч різних даних, що було детально описано в попередньому пункті, але, очевидно, що далеко не всі вони необхідні для прогнозування результату гри.

Моделі прогнозування результату футбольного матчу зводиться до пошуку функціонального відображення вигляду (формула 2.1):

$$X = \{x_1, x_2, \ldots, x_n\} \rightarrow D \in \{d_1, d_2, d_3\}, \qquad (2.1)$$

де X – вектор факторів, що впливають,

n – кількість факторів, що впливають,

D – результат гри, який може приймати три можливих значення,

$d_1$ – виграш першої команди (команди-господаря),

$d_2$ – нічия,

$d_3$ – виграш другої команди (команди-гостя) [10].

Ретельний аналіз предметної області дозволив визначити фактори, які є найбільш істотними для побудови моделі прогнозу. Дані для вхідного вектора X складаються з елементів двох типів: стимулятори і дестимулятори. Стимулятори – це фактори, при збільшенні яких ймовірність перемоги збільшується. Дестимулятори – фактори, при збільшенні яких ймовірність перемоги зменшується [8].

До стимулюючим факторам впливу відносяться [10]:

а) набрану кількість очок;

б) кількість забитих м'ячів;

в) кількість ударів по воротах;

г) кількість ударів в створ воріт;

д) кількість передач;

е) кількість точних передач;

ж) кількість кутових;

з) відсоток володіння м'ячем;

і) відпочинок (скільки днів команда не грала);

к) домашній матч;

л) сейви.

Серед дестімуліріющіх факторів можна виділити наступні:

а) місце в турнірній таблиці;

б) кількість гравців, які пропускають матч;

в) кількість пропущених м'ячів.

Вплив кожного з факторів у різній мірі впливає на результат зустрічі. Рівень впливу повинен бути визначений експертами футбольної області. Варто зауважити, що важливість факторів може змінюватися не тільки в залежності від команди, але і від матчу та противника. Можна було б подумати, що від відсотка володіння м'ячем залежить кількість голів: чим більше у команди м'яч, тим більше моментів для того, щоб забити його. Однак, як показує статистика, це не завжди так: матч може виграти команда, яка володіла м'ячем меншу частину гри.

Також варто враховувати поточного і попередніх супротивників, так як показники сумуються по кількох попередніх матчів. Якщо, наприклад, у команди були кілька слабких супротивників поспіль, то її шанс на виграш буде високим, але це може бути помилково, якщо поточний противник – сильна команда.

### 2.2.1 Метод заснований на предметної області

Розглянемо модель оцінки рейтингу команд з використанням кваліметріческого методу. Кваліметрія – наукова дисципліна, в рамках якої вивчаються методологія і проблематика комплексної, кількісної оцінки якості об'єктів будь-якої природи. В даному випадку проводиться рейтингова оцінка двох футбольних клубів, які беруть участь в матчі. На підставі цієї оцінки буде будуватися функціональне відображення попередніх результатів (входів моделі) в поточні (виходи), тобто прогноз підсумкового рахунку матчу.

Першим етапом алгоритму є виділення і ранжування факторів, що впливають. З перерахованих вище факторів виділимо тільки ту частину, яка вважається найбільш значною і загальнодоступною. Для успішного прогнозу нам потрібні дані не тільки за останній зіграний матч, але і за кілька попередніх. Також варто врахувати, що для одних значень переважно брати середньоарифметичне, для інших суму за останній матчі, а для третіх просто їх значення. Це обумовлено особливостями ваги факторів і їх природою.

Очевидно, що чинники прогнозування не є рівнозначними між собою. Тому доцільно для кожного з них вказати ваговий коефіцієнт (ступінь важливості): чим більше число, тим сильніше показник впливає на результат. Вагові коефіцієнти встановлюються емпіричним шляхом і переважно, щоб це робили експерти з футболу. Вагові коефіцієнти також можуть варіюватися в залежності від команди, наприклад, деякі команди

грають виїзні матчі краще, ніж домашні і тому подібні обставини. Якщо використовувати алгоритм протягом хоча б одного сезону, то коефіцієнти можна підібрати. Правда після кожного сезону в командах зазвичай відбуваються зміни: в складі гравців, в тренерському складі і необхідно враховувати цей факт перед новим сезоном.

У таблиці 2.1 вказані вибрані впливають фактори, тип розрахунку за останні матчі і вагові коефіцієнти. Вказані середні значення вагових коефіцієнтів.

Таблиця 2.1 – Впливаючі фактори

| Фактор | Тип розрахунку | | | Ваговий коефіцієнт |
|---|---|---|---|---|
| | Значення | Середнє | Сума | |
| Стимулюючі | | | | |
| Набрана кількість очок | + | – | Набрана кількість очок | + |
| Кількість забитих м'ячів | – | – | Кількість забитих м'ячів | – |
| Відсоток володіння м'ячем | – | + | Відсоток володіння м'ячем | – |
| Кількість ударів по воротах | – | + | Кількість ударів по воротах | – |
| Кількість ударів в створ | – | + | Кількість ударів в створ | – |
| Кількість кутових | – | + | Кількість кутових | – |
| Сейви | – | – | Сейви | – |
| Дестимулюючі | | | | |
| Кількість пропущених голів | – | – | Кількість пропущених голів | – |
| Місце в таблиці | + | – | Місце в таблиці | + |

Показники з типом розрахунку «Середнє» та «Сума» необхідно брати за певний проміжок, який обумовлений такими факторами:

а) кількість днів з останнього матчу;

б) кількість днів з останнього зміни в основному складі;

в) тип проведеного змагання і рівень попередніх суперників.

В середньому статистика береться за останні 5 матчів.

Розрахуємо відносну вагу фактору для кожної команди. Для цього скористаємося формулою 2.2 для показників-стимуляторів і формулою 2.3 для показників-дестимулятори:

$$\mathrm{Xstand_i} = \frac{\mathrm{X_i}}{\sum_{j=1}^{2} \mathrm{X_j}}, \tag{2.2}$$

$$\mathrm{Xstand_i} = 1 - \frac{\mathrm{X_i}}{\sum_{j=1}^{2} \mathrm{X_j}}, \tag{2.3}$$

де X – початкове значення показника;

i – команда;

Xstand – відносне значення показника.

З урахуванням обраних ваг для кожного коефіцієнта, розрахуємо рейтинг команд за формулою 2.4.

$$\mathrm{Ra_i} = \mathrm{Mstand_i} \cdot \mathrm{W_i}, \tag{2.4}$$

де W – матриця вагових коефіцієнтів;

Mstand – матриця, що містить відносні значення показників Xstand;

i – команда;

Ra – рейтинг команди.

Для зручності порівняння зробимо нормування рейтингу команд на одиницю скориставшись формулою 2.5:

$$\mathrm{Rstand_i} = \frac{\mathrm{Ra_i}}{\sum_{j=1}^{2} \mathrm{Ra_j}}, \tag{2.5}$$

де Ra – абсолютний рейтинг команди;

i – команда;

Rstand – рейтинг команди, нормований на одиницю.

Для інтерпретації отриманого рейтингу команд необхідно ввести так звану лінгвістичну інтервальна шкалу. У кваліметрії шкала вимірювань є засобом адекватного зіставлення і визначення чисельних значень окремих властивостей і якостей відмінності об'єктів. У нашій моделі будемо використовувати трирівневу лінгвістичну шкалу, наведену в таблиці 2.2 [11].

Таблиця 2.2 – Лінгвістична інтервальна шкала для моделі

| Рейтинг команди | Результат | Значення |
|---|---|---|
| 1 | 2 | 3 |
| Rstand ≥ 65% | Перемога | $d_1$ |
| 35% ≤ Rstand < 65% | Нічия | $d_2$ |
| Rstand < 35% | Поразка | $d_3$ |

Розроблену математичну модель можна використовувати для прогнозування результату футбольного матчу. Однак варто враховувати, що даний підхід має на увазі настроювання параметрів моделі аналітиком безпосередньо перед кожним матчем. Враховуючи факт, що кожного тижня відбувається близько сотні матчів відомих футбольних ліг, майже неможливо підтримувати значення параметрів моделі актуальними.

### 2.2.2 Метод часових рядів

Прогнозування на основі часових рядів передбачає наявність деякої числової послідовності, яка фіксує зміна прогнозованої величини в часі. Прогнозування футбольних матчів не має на увазі обробку інформації за тривалий період, так як рівень команди може істотно змінитися за короткий проміжок часу (сезон). Отже, і довгостроковий прогноз не буде володіти високою точністю. Тому виправдано для прогнозу використовувати 8-10 останніх матчів команди і будувати прогноз не більше ніж на 2 майбутні ігри [11].

Розглянемо два тимчасових ряди R1 і R2 (формула 2.6) для команд A і B, які відображає результати матчів обох команд, за обмежений проміжок часу:

$$\begin{aligned} R1 &= \{r_{1,1}, r_{1,2} \dots r_{1,n}\}, \\ R2 &= \{r_{2,1}, r_{2,2} \dots r_{2,n}\}, \end{aligned} \tag{2.6}$$

де r – результат матчу (1 – перемога, 0 – нічия, –1 – поразка);

n – кількість матчів.

Кількість матчів необхідних для прогнозування встановлюються дослідним шляхом. Ряд побудований таким чином, що результати впорядковані за зростанням, тобто $r_{1,1}$ – найдавніший матч команди A. Тоді результат n + 1 матчу для можна розрахувати за формулою 2.7:

$$r_{j,n+1} = \sum_{i=1}^{n} r_{j,i} \cdot w_i, \tag{2.7}$$

де j – номер команди (1 – команда A, 2 – команда B);

w – коефіцієнт впливу.

Вектор коефіцієнтів впливу w розраховується за формулою 2.8:

$$\begin{cases} \sum_{i=1}^{n} w_i = 1, \\ w_{k-1} \le w_k, 1 \le k \le n. \end{cases} \quad (2.8)$$

Є кілька варіантів вибору коефіцієнтів впливу. Найпростіший варіант – це вважати, що всі попередні результати рівносильно відносяться на майбутню гру. В такому випадку елементи вектора розраховуються наступним чином (формула 2.9):

$$w_i = \frac{1}{n}, 1 \le i \le n. \quad (2.9)$$

Однак, на підставі аналізу предметної області можна зробити висновок, що при такому розрахунку коефіцієнтів точність прогнозів буде низькою, оскільки найближчі матчі мають більше впливу ніж віддалені.

Так як факт того, що попередні ігри впливають на прогноз по-різному встановлено, то має сенс розглянути варіант розрахунку коефіцієнтів за допомогою арифметичної прогресії. Причому для найвіддаленішого результату використовується найменший член арифметичної прогресії. Для такого методу розрахунок необхідно буде вказати значення максимального або мінімального коефіцієнта. При прогнозуванні на підставі 8-10 результатів, рекомендуються значення не більше 0,05 для мінімального коефіцієнта і 0,5 – для максимального. Розрахунок коефіцієнтів впливу матиме різний вигляд для максимального (формула 2.10) і мінімального (формула 2.11) елемента прогресії:

$$w_i = w_n - \frac{2(w_n - 1)}{i}, 1 < i \le n, \quad (2.10)$$

$$w_i = \frac{2 - w_1 i}{i}, 1 < i \leq n, \quad (2.11)$$

де $w_1$ – максимальний коефіцієнт впливу;

$w_n$ – мінімальний коефіцієнт впливу.

Крім розрахункових коефіцієнтів впливу можна також використовувати значення, отримані емпіричним шляхом. Подібно до коефіцієнтів факторів, що впливають вони можуть варіюватися в залежності від команди і матчу. Варто взяти до уваги, що значення ваг не повинно бути більше 0,5, а краще, щоб не більше 0,3, тоді прогноз буде більш виправданим. Значення на останні кілька ігор обґрунтовано взяти близькими один до одного, а на подальші ігри поступово зменшувати.

Після вибору методу розрахунку коефіцієнта і обчислення значень результатів n + 1 для обох команд, необхідно їх співвіднести їх між собою і прийняти рішення про підсумковий результат. Розглянемо різницю (формула 2.12) між значеннями розрахованих результатів n + 1 матчу для двох команд:

$$rm = r_{1,n} - r_{2,n}. \quad (2.12)$$

Очевидно, що це різниця може приймати будь-яке значення в діапазоні [-1; 1]. Побудуємо ще одну трирівневу лінгвістичну шкалу (таблиця 2.3) для визначення результату матчу між командами A і B.

Таблиця 2.3 – Лінгвістична інтегральна шкала для ряду

| Модуль різниці | Результат | Значення |
|---|---|---|
| rm > 0,25 | Перемога команди *A* | $d_1$ |
| –0,25 ≤ Rstand < 0,25 | Нічия | $d_2$ |
| Rstand < –0,25 | Поразка команди *A* | $d_3$ |

Недоліком такого підходу є те, що в методі ніяким чином не враховується складність попередніх ігор команд, тому можна припустити, що точність такого прогнозу не буде висока.

### 2.2.3 Прогнозування за допомогою ІНМ

Імовірнісна нейронна мережа є мережею з прямою передачею даних або ж мережею прямого поширення. Мережі прямого поширення допускають проходження сигналу тільки в одному напрямку – від входів до виходів [5]. Структурно елементи мережі можна розділити на чотири складові: входи X, функції нейронів активації Ф, суматори та вихідний шар (рисунок 2.1).

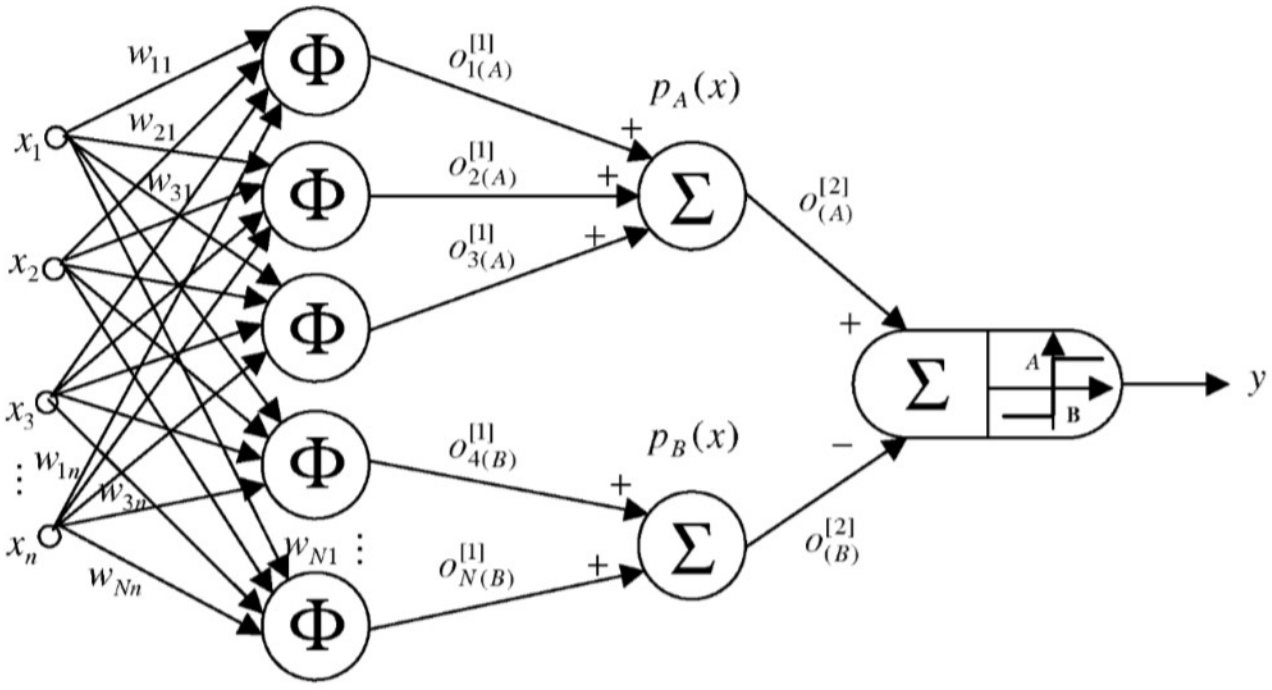


Рисунок 2.1 – Структурна схема ІНМ

Вихідною інформацією для створення мережі є навчальна вибірка образів, утворена набором або пакетом n-мірних векторів x(1), x(2), ..., x(N) з відомою класифікацією, причому місце конкретного образу в пакеті

значення не має. Передбачається також, що $N_A$ векторів відносяться до класу A, $N_B$ – до класу B, тобто:

$$N_A + N_B = N. \tag{2.13}$$

Кількість нейронів у шарі образів одно N (по одному нейрону на кожен навчальний образ), а їх синаптичні ваги визначаються значеннями цих образів так, що (формула 2.15):

$$w_{ji} = x_i(j), i = 1, 2, \dots, n; j = 1, 2, \dots N. \tag{2.14}$$

Очевидно, що навчання в даному випадку зводиться до одноразової установки ваг, що робить його надзвичайно простим. Кожен з нейронів шару образів обчислює зважену суму вхідних сигналів і перетворює її за допомогою нелінійної активаційної функції так, що на виході нейронів першого прихованого шару з'являється сигнал (формула 2.15):

$$o_j^1 = \Phi(\|x(k) - w_j(k)\|, \sigma). \tag{2.15}$$

Шар підсумовування утворений елементарними суматорами для класів A та B (в загальному випадку по одному на кожен клас), які просто підсумовують (формула 2.16) виходи нейронів прошарку образів:

$$\begin{cases} o_j^2(k) = \sum_{j=1(A)}^{N_A(A)} o_j^1(k), \\ o_j^2(k) = \sum_{j=(N_A(A)+1)B}^{N_B(B)} o_j^1(k). \end{cases} \tag{2.16}$$

Суми і є парзеновскими оцінками невідомих щільності ймовірностей $p_A(x)$ і $p_B(x)$. У вихідному нейроні мережі, що реалізує по суті елементарну операцію порівняння обчислених значень $p_A(x(k))$ і $p_B(x(k))$, визначається приналежність висунутого способу x(k) класу A або B. Клас, для якого обчислена функція буде мінімальна є переможцем. У класичному варіанті існує зворотна залежність.

Головним достоїнством імовірнісних нейронних мереж є простота проектування і навчання. Основний їх недолік визначається різким зростанням числа нейронів в першому прихованому шарі при великому обсязі навчальної вибірки [6].

## 2.3 Вибірка футбольної статистки

З розвитком мережі інтернет дані майже будь-які дані можна знайти за лічені секунди. Проте дані по своїй сутності являються неструктурованим та неповними. Перш ніж починати розробку моделі прогнозування необхідно перетворити дані на інформацію. Це дозволить отримати прогноз – знання, що являються останнім етапом еволюції даних.

На сайтах присвячених футбольним змаганням зазвичай на одній сторінці можливо подивитися результати у двох видах: загальну інформацію поточного туру (рисунок 2.2) та детальну інформацію однієї гри (рисунок 2.3). Загальна інформація відповідає таблиці 1.1, а інформація стосовно гри – рисунку 1.4. Кожен вид відображення інформації корисний. Проте для поставленої цілі, а саме: прогнозування за допомогою ІНМ, другий тип відображення є більш цінний, бо включає в себе більше детальної інформації щодо матчу.

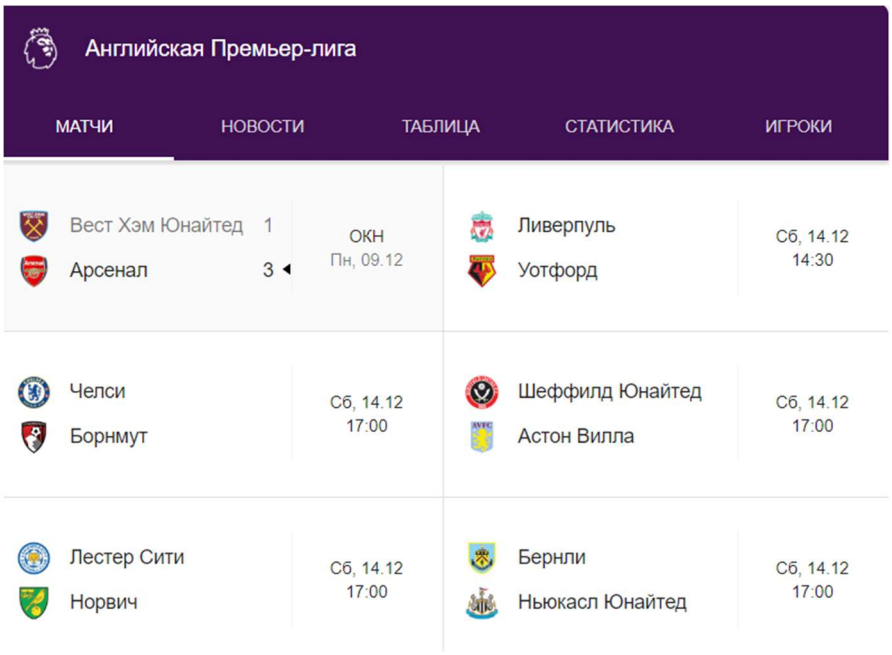


Рисунок 2.2 – Інформація по туру

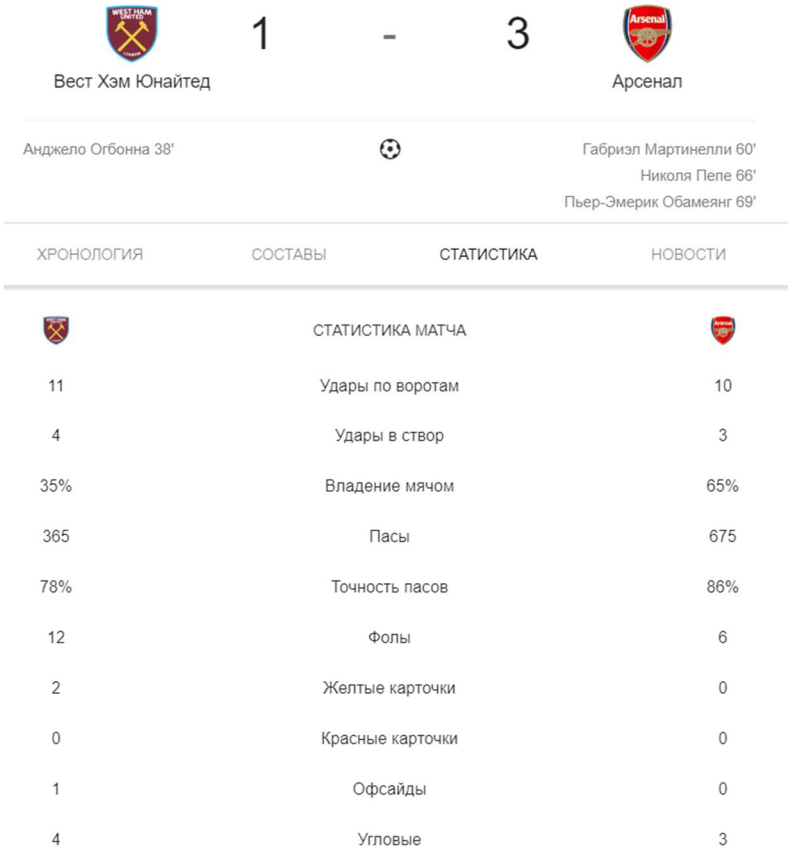


Рисунок 2.3 – Інформація по грі

Більшість груп, які займаються створенням прогнозів результатів футбольних матчів мають власну БД, яка містить відомості по зіграним матчам. Вона може бути поверхневою, тобто містити дані тільки по матчам, та поглибленою – містити статистику по гравцям. Збір поглибленої статистики клопітливий процес, який потребую великої кількості ресурсів, як обчислювальних, так і людських. Деякі ресурси розміщують дані у вільний доступ, або створюють API для звернення до їхніх серверів. Серед найбільш відомих можна виділити наступні:

а) football.db (дані у форматі TXT);

б) football-data.co.uk (дані у форматі CVS):

в) api-football (дані отримуються через API);

г) football-data.org (дані отримуються через API).

Football.db має значну кількість матчів, проте інформація обмежена – лише результат гри. Дані на football-data.co.uk відображують доволі детальну статистичну інформацію про матчі багатьох турнірних ліг за останні два десятиріччя. Два останні сервіси є потоковими та мають преміум аккаунти. Під час безкоштовного використання вони мають обмежують кількість запитів на день/місяць.

Найкращім варіантом для навчальної вибірки ІНМ є дані, які мають найбільше кількісних характеристик гри. Отже, для прогнозування обрані дані з третього сервісу. Дані мають наступну структуру:

а) Date – дата гри;

б) HomeTeam – команда-господар (H);

в) AwayTeam – команда-гість (A);

г) FTHG – кількість м’ячів забитих H;

д) FTAG – кількість м’ячів забитих A;

е) FTR – результат гри;

ж) HS – удари по воротах H;

з) AS – удари по воротах A;

и) HST – удари в стійку воріт H;

і) AST – удари в стійку воріт А;

к) HC – кутові Н;

л) AC – кутові А;

м) HF – фоли Н;

н) AF – фоли А;

о) HY – жовті картки Н;

п) AY – жовті картки А;

р) HR – червоні картки Н;

с) AR – червоні картки А.

Такої кількості параметрів повинно бути достатньо для того, щоб нейронна мережа змогла проаналізувати існуючі залежності і створити прогноз достатньої точності.

# 3 ПРОГРАМНА РЕАЛІЗАЦІЯ

## 3.1 Вибір засобів розробки

Для реалізації і дослідження ефективності обрано середовище розробки Matlab, так вона має зручні засоби для програмування використовуваних алгоритмів і побудови графіків залежностей.

Мова Matlab – високорівнева інтерпретуєма мова програмування, що включає засновані на матрицях структури даних, широкий спектр функцій, інтегроване середовище розробки, об'єктно-орієнтовані можливості і інтерфейси до програм, написаним на інших мовах програмування. Остання характеристика є важливою при виборі в даній ситуації, бо розробляємий модуль в подальшому буде вбудовуватися в веб-сервіси або прикладну програму.

До переваг Matlab слід віднести:

а) масштабність: велика кількість математичних, фізичних, статистичних функцій, оптимізовані обчислення, великі графічні можливості;

б) безвідмовність: Matlab має об’ємну документацію, яка покриває усі засоби програмування;

в) прозорість: вихідний код всіх засобів є відкритим, з ним можна ознайомитися і, при необхідності, внести в нього корективи;

г) універсальність: Matlab широко використовується для побудови складних обчислювальних моделей різних галузей [14].

Deep Learning Toolbox (раніше Neural Network Toolbox) – це один із пакетів середовища Matlab, що містить засоби для проектування, моделювання, розробки та візуалізації нейронних мереж.

Пакет дає можливість реалізації різноманітних нейромережевих підходів і має відкритий вихідний код. Пакет містить програмні функції для створення скриптів і графічний інтерфейс користувача для швидкого

створення мереж без написання програмного коду. Серед головних переваг прикладного програмного модуля можна виділити наступне:

а) графічний інтерфейс користувача для покрокового створення, навчання і імітаційне моделювання нейронних мереж;

б) підтримка найбільш поширених керованих і некерованих мережевих структур;

в) повний перелік навчальних функцій для тестування;

г) динамічні алгоритми навчання мереж, що включають тимчасову затримку, нелінійну авторегресії (NARX), ланцюгові і настроюються динамічні структури;

д) блоки Simulink для створення нейронних мереж і розвинених блоків для систем контролю;

е) автоматична генерація блоків Simulink з об'єктів нейронної мережі;

ж) модульне подання мережі, що дозволяє створювати необмежену кількість вхідних шарів і об'єднаних мереж, а також графічне представлення архітектури мережі;

з) функції попередньої і постобробки і блоки Simulink для поліпшення процесу навчання і оцінки продуктивності мережі;

і) візуалізація топології і процесу навчання нейронної мережі [13].

## 3.2 Препроцесінг вхідного образа

Для прогнозування ІНМ необхідно знати поточні значення вектора-образа X. Однак, очевидно, що на момент прогнозування, тобто до того, як відбулась гра, ми не маємо цих даних. Для вирішення цієї проблеми пропонується два методи.

Перший базується на окремих результатах попередніх матчів. Для кожної команди створюється набір даних з її історичних матчів. На основі характеристик якості гри команди у минулих матчах розраховується

передбачуване значення цих показників для поточної гри. Розрахувати значення можна декількома способами:

а) середнє значення;

б) середнє зважене значення (по «давнині» гри або рейтингу команди);

в) мода або медіана.

У таблиці 3.1 можна побачити помилку при розрахунку методами середнього та середньозваженого значення.

Таблиця 3.1. Помилка розрахунку першого метода

| Спосіб | HS | HST | HF | HC | HY | HR |
|---|---|---|---|---|---|---|
| Середнє | 6,7059 | 0,4706 | 0,3529 | 4,0588 | -0,4118 | 0,0588 |
| Середнє зважене | 11,1843 | 0,2925 | -0,5409 | 3,9939 | -0,2792 | 0,0020 |

Також при створенні масиву матчів можна спиратися на фактори домашньої/гостьової гри. Наприклад для команди-господаря, обирати тільки домашні, а для команді-гостя навпаки – тільки виїзді. Також можна комбінувати набори матчів.

Другий підхід базується на кореляційнозв’язанних значеннях. Для пошуку вхідного вектора-образа будуть враховуватися тільки матчі проти однакових команд.

Нехай необхідно спрогнозувати результат матчу A проти B, причому A – команда-господар, а B – команда-гість. Тоді виберемо з навчальної вибірки такі пари матчів, в яких команда A була господарем поля, а команда B грала на виїзді, та мають однакового суперника C. Далі необхідно привести показники команди C до однакових значень в обох матчах та врахувати це в показниках команд A і B. Отримані після перерахунку значення для команд можна використовувати в якості вхідного вектора для нейронної мережі. У таблиці 3.2 можна побачити помилку при розрахунку даним методом.

Таблиця 3.2 – Помилка розрахунку другого метода

| HS | HST | HF | HC | HY | HR |
|---|---|---|---|---|---|
| 4,1968 | -0,2389 | -0,1574 | 3,1398 | 0,5267 | -0,1025 |

Таким чином отримані значення будуть відображати більш достовірні характеристики домашньої та виїзної гри для команд A та B. Однак у такому підході існує недолік. Статистика показує, що зазвичай команди показують кращі результати в домашніх матчах. А при пошуку спільних матчів команда С грає на різних сторонах.

Для того, що нівелювати отриману похибку також використовувати для прогнозу інші комбінації домашніх/гостьових матчів. Проте необхідно зазначити, що матчі, де команди А та В попарно змінюють цю характеристику, теоретично мають створювати найгірший прогноз, бо спотворюють фізичний зміст кореляції матчів.

На виході описаного алгоритму отримуємо набір вхідних векторів-образів для ІНМ. На цьому етапі препроцесінгу можна отримати підсумковий прогноз двома способами. Перший – це застосувати алгоритм пошуку середніх значень. В другому для кожного отриманого вектора отримується прогноз, а після цього результатом матчу стає той, який настав для найбільшої кількості вхідних векторів. Якщо два результати мають рівні кількості векторів, то можна обирати випадково, або по останньому матчу.

## 3.3 Реалізація алгоритму прогнозування

Після повного аналізу та розуміння методу вирішення задачі прогнозування переходжу до програмної реалізації. Як вже було сказано, обрано середовище Matlab. У таблиці 3.3 представлено перелік вбудованих функцій Matlab та їх коротка характеристика, які були використані для створення програмного модуля.

Таблиця 3.3 – Функцій Matlab

| Функція | Опис |
|---|---|
| clear | Видаляє задані змінні з робочого простору |
| find | Шукає індекс елемента у векторі по умовам |
| height | Повертає кількість строк таблиці |
| isempty | Перевіряє значення на пустоту |
| ismember | Перевіряє належність строки до таблиці |
| max | Повертає максимальний елемент вектора |
| newpnn | Створює імовірнісну нейронну мережу |
| randi | Генерує цілу випадкову величину в заданому діапазоні |
| readtable | Імпортує файл CSV до робочого простору |
| size | Повертає розмір матриці |
| string | Перетворює значення до строкового типу |
| struct2table | Перетворює структуру в таблицю |
| sum | Підсумовує значення вектора |
| zeros | Формує масив заповнений нулями заданого розміру |

Виходячи з аналізу видів препроцесінгу, для програмної реалізації було обрано той, який дає найближчі результати до реальних, тобто другий. Спершу необхідно імпортувати дані. Це можливо зробити використовуючи функцію, або скориставшись клавішою інтерфейсу. При використані розробленого модуля сторонньою програмою можливо передавати навчальну вибірку у якості параметру. Під час розробки та тестування систем класифікації зазвичай навчальну вибірку перемішують. У випадку з футбольними матчам це недопустимо, бо неможна допустити впливу майбутніх матчів на поточний.

Далі необхідно розробити функцію препроцесінгу вхідного вектора. Для тестування я обрала другий спосіб. Для цього в окремому m-файл я зробила функцію (таблиця 3.4). Функція має три вхідних параметри:

результати домашнього матчу однієї команди, результати виїзного матчу другої команди та режим функціонування (вказує на те, яким чином були отриманні дані). Повертає функція вектор образ.

Таблиця 3.4 – Прототип функції препроцессінгу

| function predict = getPredictByTwoGame (HomeTeamGame, AwayTeamGame, mode) |
|---|

Далі необхідно розробити функцію пошуку кореляційних матчів для команд. Завдяки вбудованим функціям пошук необхідних матчів в змінній типу таблиця зробити легко (таблиця 3.5). Команда шукає усі матчі, для яких команда-господар – Man United. Результат відображено на рисунку 3.1.

Таблиця 3.5 – Пошук матчів команди

| footballdata(ismember(footballdata(1:gameStatsPredictedNo-1,:).HomeTeam, 'Man United'),:) |
|---|

| Date | HomeTeam | AwayTeam | FTHG | FTAG | FTR | HS | AS | HST | AST |
|---|---|---|---|---|---|---|---|---|---|
| 10/08/2018 | 'Man United' | 'Leicester' | 2 | 1 | 'H' | 8 | 13 | 6 | 4 |
| 27/08/2018 | 'Man United' | 'Tottenham' | 0 | 3 | 'A' | 23 | 9 | 5 | 5 |
| 22/09/2018 | 'Man United' | 'Wolves' | 1 | 1 | 'D' | 15 | 11 | 6 | 8 |
| 06/10/2018 | 'Man United' | 'Newcastle' | 3 | 2 | 'H' | 18 | 13 | 10 | 8 |
| 28/10/2018 | 'Man United' | 'Everton' | 2 | 1 | 'H' | 14 | 14 | 10 | 6 |
| 24/11/2018 | 'Man United' | 'Crystal Palace' | 0 | 0 | 'D' | 12 | 13 | 5 | 2 |
| 05/12/2018 | 'Man United' | 'Arsenal' | 2 | 2 | 'D' | 10 | 9 | 7 | 4 |
| 08/12/2018 | 'Man United' | 'Fulham' | 4 | 1 | 'H' | 20 | 10 | 11 | 4 |
| 26/12/2018 | 'Man United' | 'Huddersfield' | 3 | 1 | 'H' | 16 | 10 | 10 | 2 |
| 30/12/2018 | 'Man United' | 'Bournemouth' | 4 | 1 | 'H' | 11 | 7 | 8 | 3 |
| 19/01/2019 | 'Man United' | 'Brighton' | 2 | 1 | 'H' | 20 | 7 | 5 | 3 |
| 29/01/2019 | 'Man United' | 'Burnley' | 2 | 2 | 'D' | 28 | 6 | 9 | 4 |
| 24/02/2019 | 'Man United' | 'Liverpool' | 0 | 0 | 'D' | 6 | 7 | 3 | 1 |
| 02/03/2019 | 'Man United' | 'Southampton' | 3 | 2 | 'H' | 16 | 6 | 6 | 3 |
| 30/03/2019 | 'Man United' | 'Watford' | 2 | 1 | 'H' | 8 | 20 | 5 | 8 |
| 13/04/2019 | 'Man United' | 'West Ham' | 2 | 1 | 'H' | 14 | 18 | 4 | 4 |
| 24/04/2019 | 'Man United' | 'Man City' | 0 | 2 | 'A' | 12 | 8 | 1 | 5 |
| 28/04/2019 | 'Man United' | 'Chelsea' | 1 | 1 | 'D' | 7 | 16 | 5 | 3 |
| 12/05/2019 | 'Man United' | 'Cardiff' | 0 | 2 | 'A' | 26 | 13 | 10 | 4 |

Рисунок 3.1 – Результати пошуку матчів

Таким чином можна відбирати матчі, накладаючи умови на будь-яку характеристику. У роботі таблиця обмежувалась тільки по найменуванню команди, однак можна створювати и більш складні умови відбору, комбінуючи декілька параметрів. Це може бути корисно для пошуку прихованих закономірностей.

Наступним етапом була розробка різних функцій пошуку прогнозу для комбінацій, а саме: господар-гість (НА), господар-господар (НН), гість-господар (АН). Завдяки розробці окремих модулів для кожного типу, в подальшому використані можливо буде комбінувати матчі в будь-якій формі. Алгоритм пошуку прогнозу містить наступні структурні елементи:

а) пошук матчів команди-господаря;

б) пошук матчів команди-гостя;

в) пошук їх спільних суперників при різних умовах (НА, НН, АН):

г) препроцесінг отриманих спільних матчів;

д) отримання прогнозу по всім матчам;

е) агрегація отриманих прогнозів – отримання підсумкового прогнозу.

Кожен розробник знає, що якщо програма відпрацювала без помилок з першого разу и повернула якийсь результат, то швидше за все вона працює неправильно. І це найскладніше всього вирішити, необхідно ретельно відслідковувати значення кожної змінної на кожному кроці. Саме з такою проблемою я зіткнулась на цьому етапі розробки. А все через відсутність у Matlab так званого збирача сміття і існування робочого простору на рівні сеансу. Це створює необхідність контролювати використання змінних та при необхідності вручну очищати їх значення.

Для більш всебічного аналізу використання імовірнісної нейронної мережі для прогнозування результатів футбольних матчів розробила подібний до першого алгоритм, але з трохи іншим порядком дій:

а) пошук матчів команди-господаря;

б) пошук матчів команди-гостя;

в) пошук їх спільних суперників при різних умовах (НА, НН, АН):

г) препроцесінг отриманих спільних матчів;

д) агрегація всіх спільних матчів;

е) отримання підсумкового прогнозу по агрегованим даним.

Відмінність від першого підходу полягає в зміні порядку агрегації та прогнозування. Якщо в першому випадку для всіх результатів отриманих зі спільних матчів отримуються прогнози, а після цього за допомогою агрегації відбувається вибір остаточного результату, то в цьому випадку навпаки. Спочатку знаходиться єдиний вектор штучних показників методом середнього значення, а потім отриманий вектор використовується для прогнозування. Такий підхід має меншу асимптотичну складність, бо для кожного матча необхідно тільки один раз зробити пошук за допомогою ІНМ, тоді як для першого кількість пошуків дорівнює кількості спільних матчів.

Останнім кроком розробки була підготовка до тестування. Для цього розроблено спеціальний модуль, який циклічно викликає різні комбінації пошуку матчів, збирає інформацію щодо помилок прогнозування та відображає отримані результати графічно.

# 4 АНАЛІЗ РЕЗУЛЬТАТІВ ТЕСТУВАННЯ

Останній етап виконання роботи – це проведення експериментального дослідження. В результаті цього етапу будуть отримані дані про точність реалізованих алгоритмів прогнозування футбольних матчів і зроблено висновок про можливість їх застосування для предметної області.

Для проведення експериментів будуть використовуватися результати сезону 18/19 англійської Прем'єр-Ліги, результати яких вже відомо. Це дасть можливість порівняти прогнозований результат з реальним, тим самим оцінити точність алгоритмів.

Запустимо розроблені алгоритм на різних розмірах навчальної вибірки. Розмір навчальної вибірки впливає на кількість нейронів у мережі, а препроцесінг відбувається на всіх спільних іграх цього сезону.

Слід також відмітити, що з ростом вибірки для препроцесінгу збільшується час отримання прогнозу. Це пов’язано з тим, що в цьому методі отримання підсумкового прогнозу спочатку знаходяться прогнози для всіх спільних, а тільки після агрегації отримуємо результат. У таблицях 4.1 та 4.2 відображені результати тестування – кількість помилок прогнозування  та точність прогнозу.

Таблиця 4.1 – Кількість помилок

| Типи матчів | Розмір навчальної вибірки | | | | |
|---|---|---|---|---|---|
| | 15 | 30 | 50 | 100 | 150 |
| НН | 178 | 174 | 174 | 158 | 126 |
| НА | 158 | 154 | 155 | 138 | 118 |
| АН | 209 | 214 | 216 | 193 | 159 |
| НН + НА | 169 | 169 | 161 | 146 | 117 |
| Усі | 185 | 185 | 185 | 184 | 184 |

Таблиця 4.2 – Точність прогнозу

| Типи матчів | Розмір навчальної вибірки | | | | |
|---|---|---|---|---|---|
| | 15 | 30 | 50 | 100 | 150 |
| НН | 0,5123 | 0,5029 | 0,4727 | 0,4357 | 0,4522 |
| НА | 0,5671 | 0,5600 | 0,5303 | 0,5071 | 0,4870 |
| АН | 0,4274 | 0,3886 | 0,3455 | 0,3107 | 0,3087 |
| НН + НА | 0,5370 | 0,5171 | 0,5121 | 0,4786 | 0,4913 |
| Усі | 0,4932 | 0,4714 | 0,4394 | 0,3429 | 0,2000 |

При теоретичному дослідженні була висунута гіпотеза про те, що при попарній зміні домашніх/гостьових матчів, спотвориться сенс методу та прогноз буде найгіршим. Як можна побачити в таблиця результати для АН найгірші, що підтверджує висунуту гіпотезу. В подальших тестування цей тип матчів розглядатися не буде.

Для наглядного відображення зміни точності побудовано графік залежності значення точності від розміру навчальної вибірки (рисунок 4.1). Проаналізувавши дані можна зробити наступні висновки:

а) при використанні для прогнозу домашніх матчів команди-господаря і виїзних матчів для команди-гостя результати прогнозу є найбільш природним для цього методу і точність, яку він забезпечує – найліпша майже для будь-якого об’єму навчальної вибірки;

б) зростання навчальної вибірки, а отже зростання кількості нейронів у шарі образів зменшує точність прогнозування для всіх способів відбору матчів для препроцесінгу.

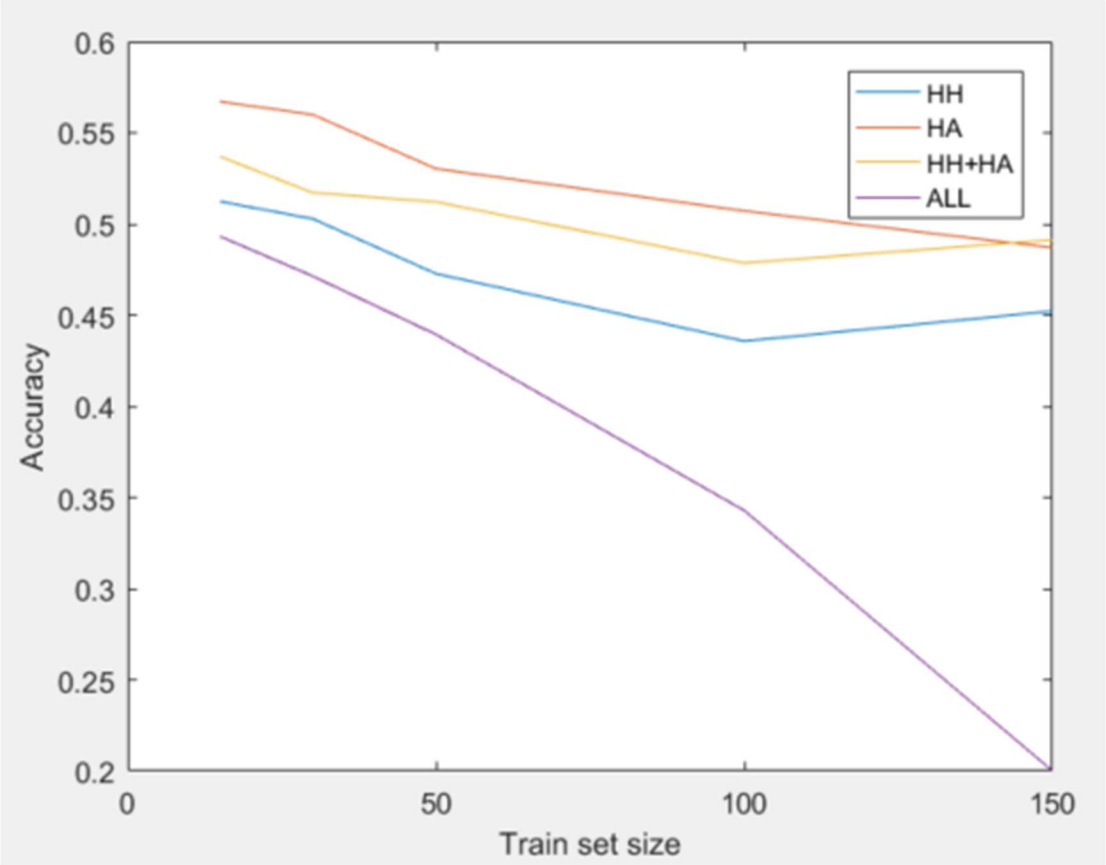


Рисунок 4.1 – Точності першого метода для різних типів матчів

Після дослідження даних отриманих за допомогою препроцесінгу, має сенс перевірити, які результати дадуть справжні дані матчів, якщо протестувати їх на розробленій мережі. Виявилося, що на справжніх даних матча обрана нейронна мережа дає гірші результати, чим при розрахунку за допомогою будь-яких типів матчів.

Однак, більш цікаво те, що існує залежність між точністю прогнозу по реальним даним і при отриманні коефіцієнтів по всім типам матчів (який показав найгірше значення точності). На рисунку 4.2 графічно відображені залежності точності для найкращої точності прогнозування, найгіршої та точність реальних даних. Можна побачити, що прогнозування на штучно отриманих даних дає суттєво кращий результат, чим на справжніх даних.

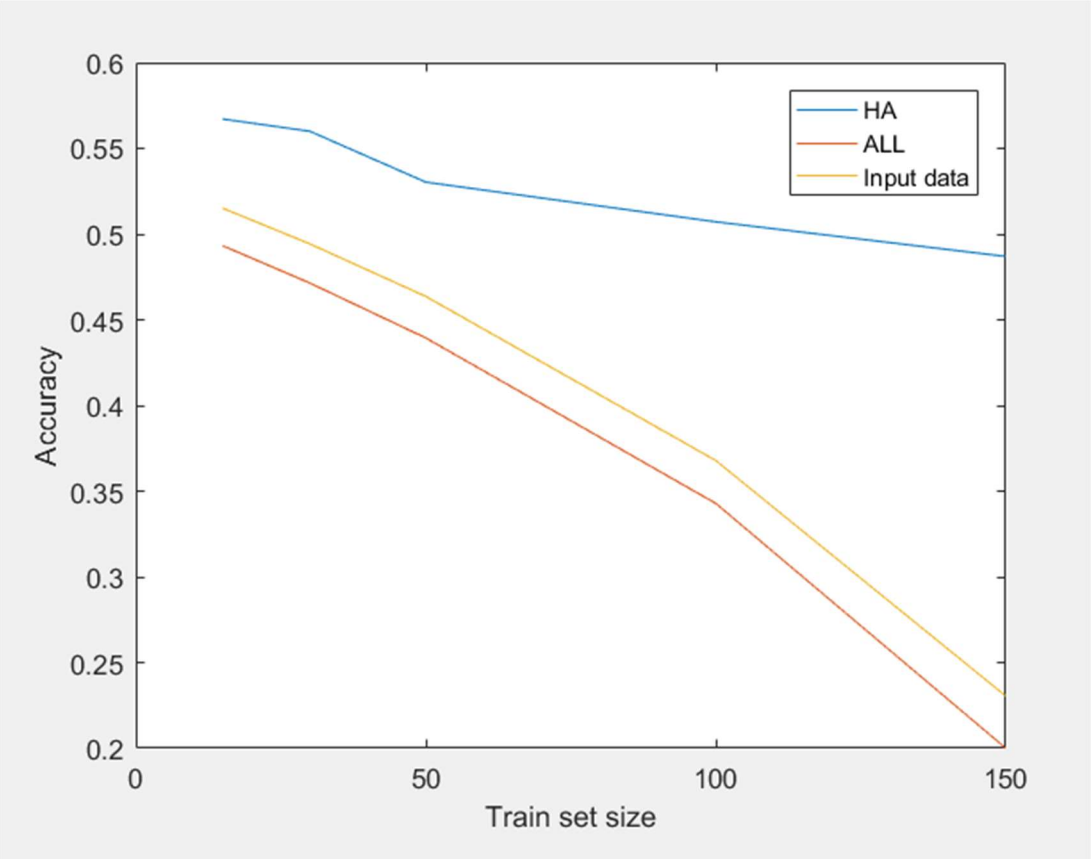


Рисунок 4.2 – Порівняння першого метода з справжніми даними

З отриманих результатів можна зробити висновок, що імовірнісна нейрона мережа не спроможна дати стерпні прогнози на тих показниках матчів, які були обрані в роботі. Проте той факт, що штучні дані, на мою думку, дають достатньо високу точність для того, щоб розвивати подальше використання імовірнісної нейронної мережі для прогнозування результатів футбольних матчів.

Далі протестовано другий спосіб формування прогнозу. Для того, щоб мати можливість порівняти два способи набір даних, обсяги тестових навчальних вибірок та спосіб відображення результатів залишимо без змін. Результати тестування другого способу у таблицях 4.3 та 4.4.

Таблиця 4.3 – Кількість помилок

| Типи матчів | Розмір навчальної вибірки | | | | |
|---|---|---|---|---|---|
| | 15 | 30 | 50 | 100 | 150 |
| НН | 178 | 186 | 189 | 158 | 148 |
| НА | 160 | 177 | 176 | 160 | 145 |
| АН | 220 | 227 | 225 | 176 | 145 |
| НН + НА | 181 | 192 | 181 | 156 | 150 |
| Усі | 216 | 214 | 215 | 199 | 211 |

Таблиця 4.4 – Точність прогнозу

| Типи матчів | Розмір навчальної вибірки | | | | |
|---|---|---|---|---|---|
| | 15 | 30 | 50 | 100 | 150 |
| НН | 0,5123 | 0,4686 | 0,4273 | 0,4357 | 0,3565 |
| НА | 0,5616 | 0,4943 | 0,4667 | 0,4286 | 0,3696 |
| АН | 0,3973 | 0,3514 | 0,3182 | 0,3714 | 0,3696 |
| НН + НА | 0,5041 | 0,4514 | 0,4515 | 0,4429 | 0,3478 |
| Усі | 0,4082 | 0,3886 | 0,3485 | 0,2893 | 0,0826 |

Згідно до таблиць при тестування найкращі результати точності, як і для першого способу, при пошуку спільних матчів на сторонах відповідних до сторін гри, яка прогнозується. Однак найгірші результати в цьому випадку для ігор команд на усіх сторонах.

Також можна побачити на рисунку 4.3, що усі типи, крім найгіршого, при великому обсязі навчальної вибірки мають приблизно рівну точність. А для типів НН, НА та НН + НА вони приблизно рівні при будь-якому обсязі вибірки. З цього можна зробити висновок, що даний спосіб менш чутливий до типу матчів, на основі яких будується штучний вхідний вектор-образ.

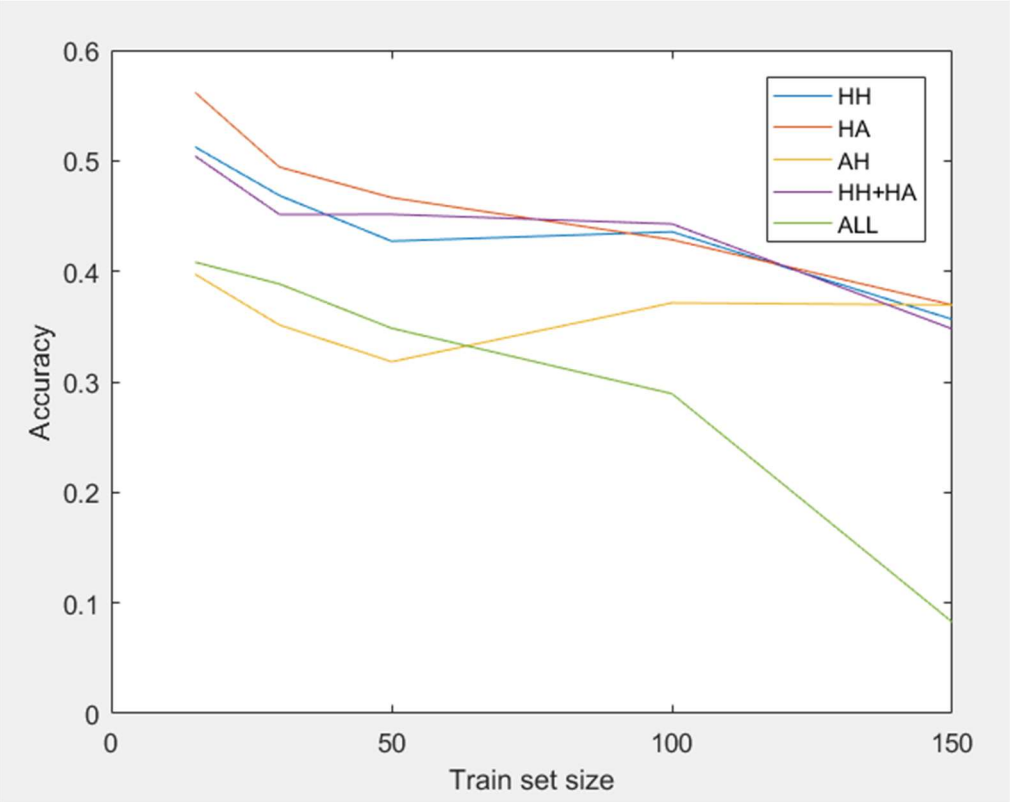


Рисунок 4.3 – Точності другого метода для різних типів матчів

Для порівняння двох розроблених у роботі методів зведемо дані по найкращим и найгіршим точностям (таблиця 4.5) та зробимо висновки:

Таблиця 4.5 – Порівняння методів отримання прогнозу

| Результат | Розмір навчальної вибірки | | | | |
|---|---|---|---|---|---|
| | 15 | 30 | 50 | 100 | 150 |
| Кращий першого метода | 0,5671 | 0,5600 | 0,5303 | 0,5071 | 0,4870 |
| Кращий другого метода | 0,5616 | 0,4943 | 0,4667 | 0,4286 | 0,3696 |
| Гірший першого метода | 0,4274 | 0,3886 | 0,3455 | 0,3107 | 0,3087 |
| Гірший другого метода | 0,4082 | 0,3886 | 0,3485 | 0,2893 | 0,0826 |

а) значення точності першого методу кращі для обох результатів;

б) перший метод більше відповідає фізичному сенсу підходу до пошуку штучного вектора через спільні матчі;

в) другий метод невеликій навчальній виборці дає достатньо хороші результати точності;

г) другий метод менш чутливий до типів матчів.

Точність розробленого модуля достатня для його подальшого розвитку та використання в майбутньому в великих комерційних прикладних програм та веб-сервісах.

# ВИСНОВКИ

Метою даної атестаційної роботи є розробка методів прогнозування результатів футбольних матчів з використанням імовірнісних нейронних мереж, оцінки їх точності і отримання висновку про можливість його застосування до футбольної галузі.

Прогнозування – це область, яка може бути застосована до будь-яких подій існуючого світу. В якості вхідних даних для прогнозування виступає набір даних обраної предметної області. Кожна подія – індивідуальна та має свій власний набір передумов. З цього випливає, що поточний стан розвитку алгоритмів не припускав отримання даних будь-якого типу на вхід для отримання прогнозу.

Футбол бере свої витоки ще в 19 столітті, його правила давно сформульовані і тим не менше це не скасовує факт, що ця ланка сучасного спорту є найпопулярнішим серед фанатів спортивних змагань. Як би там не було результат футбольного матчу – це складне завдання для прогнозування, так як результат залежить про кореляції безлічі, іноді безпосередньо не пов'язаних, особистісних і спортивних подій. Незважаючи на це сучасні алгоритми прогнозування можуть нехтувати відсутністю або неточністю наявних даних, але давати прогноз непоганий точністю.

Одним з найважливіших етапів при прогнозуванні результатів футбольних змагань з використанням ІНМ є препроцесінг. Препроцесінг створює штучний вхідний вектор-образ, від якого повністю залежить прогноз. Для його отримування можна використовувати різноманітні, які в більшості випадків базуються на моделі предметної області.

В роботі представлений спосіб, основною ідеєю якого є пошук матчів з однаковим супротивником. Відібрані матчі оброблюються и параметри отриманого результату зв’язані між собою не випадково, а реальною грою. В результаті цього підвищується точність розробленого модуля прогнозування.

Згідно до проведеного тестування прогнозування результатів футбольних змагань з використанням імовірнісна нейронна мережа, можна зробити висновок, що дана мережа може бути основою для створення системи, яка займається букмекерською діяльністю. Найкраща отримана точність дорівнює 0,5671. Зазвичай таке значення точності вважається недостатнім для того, щоб вважати метод успішним. Проте у випадку з футбольною областю, я вважаю результати достатніми, так як предметна область є складною для прогнозування через велику частку людського фактору та відсутність постійних правил, які можна було би використовувати для основи прогнозу.

# ПЕРЕЛІК ДЖЕРЕЛ ПОСИЛАНЬ

1. Total sportek. URL : https://www.totalsportek.com (дата звернення: 04.06.2018).

2. BetUnion24. URL : https://betunion24.ru (дата звернення: 04.12.2019).

3. Чучуева И.А. Модель прогнозирования временных рядов по выборке максимального подобия: дис. ... канд. техн. наук: 05.14.02 / МГТУ им. Н.Э.Баумана, 2012. 154 с.

4. Ротштейн А.П. Интеллектуальные технологии идентификации: нечеткие множества, нейронные сети, генетические алгоритмы: монография. Винница : Вінниця-УНІВЕРСУМ, 1999. 295 с.

5. Бодянский Е.В., Дейнеко А.А., Куценко Я.В. (Хаустова Я.В.). Ядерная самоорганизующаяся карта на основе радиально-базисной нейронной сети // Электротехнические и компьютерные системы. Одесса, 2015. № 20 (96). С. 97–105.

6. Бодянский Е.В., Руденко О.Г. Искусственные нейронные сети: архитектура, обучение, применения. Харьков : ТЕЛЕТЕХ, 2004. 369 с.

7. Букмекер. URL : https://ru.wikipedia.org/wiki (дата звернення: 04.12.2019).

8. Правила игры. URL : http://www.ff.spb.ru/sites/default/files/page-attachs/pravila_igry_2016-2017.pdf (дата звернення: 04.12.2019).

9. Англия. Премьер-Лига 2017/2018 результаты матчей. URL : https://www.soccer.ru/tournament/england/results (дата звернення: 04.12.2019).

10. Метод взвешенной суммы показателей для прогнозирования футбольных матчей. URL : https://bets.today/ru/articles/weighted-sum-of-indexes (дата звернення: 07.06.2018).

11. Бехтер Л.В. Прогнозирование исхода футбольных матчей // Социально-экономическое развитие АР Крым: проблемы и перспективы. 2012. №6. С. 268–270.

12. Dolores: a model that predicts football match outcomes from all over the world. URL : https://www.researchgate.net/publication/324926213_Dolores_a_model_that_predicts_football_match_outcomes_from_all_over_the_world (дата звернення: 02.12.2019).

13. Mathworks. URL : https://www.mathworks.com (дата звернення: 03.12.2019).

14. Визуальное моделирование финансовых операций в среде Simulink / Matlab / С.В. Юдин, В.Г. Степанов, Т.В. Степанова, И.И. Румянцева, И.К. Архипов, Д.И. Якушин, В.И. Абрамова // Научно-методический электронный журнал «Концепт». 2015. № S6. С. 36–40. URL : http://e-koncept.ru/2015/75104.htm (дата звернення: 05.12.2019).

15. Официальные сайты клубов Англии. URL : http://fanat.ua/links/england/official/ (дата звернення: 04.12.2019).

16. HTML Table. URL : www.w3schools.com/html/html_tables.asp (дата звернення: 05.12.2019).

17. Пирогов С.В. Социальное прогнозирования и проектирование. Москва : Проспект, 2016. 376 с.