КИЇВСЬКИЙ НАЦІОНАЛЬНИЙ УНІВЕРСИТЕТ
ІМЕНІ ТАРАСА ШЕВЧЕНКА

НАВЧАЛЬНО-НАУКОВИЙ ІНСТИТУТ ВИСОКИХ ТЕХНОЛОГІЙ

Завідувач кафедри молекулярної біотехнології та біоінформатики кандидат біологічних наук Нипорко Олексій Юрійович

Протокол №_____ засідання кафедри

від “____” ____________20___ р.

ПІДВИЩЕННЯ ЯКОСТІ РОЗПІЗНАВАННЯ БІЛКІВ З ВНУТРІШНЬОЮ НЕВПОРЯДКОВАНІСТЮ ЗА ДОПОМОГОЮ МЕТОДІВ МАШИННОГО НАВЧАННЯ

Випускна кваліфікаційна робота магістра
студента спеціальності
091 Біологія
ОП «Біоінформатика і структурна біологія»
Халецького Євгена Валентиновича
Науковий керівник
доцент кафедри молекулярної
біотехнології та біоінформатики
к.б.н. **Нипорко Олексій Юрійович**

Оцінка захисту роботи

____

Київ – 2022 р.

АНОТАЦІЯ

Халецький Є. В. Підвищення якості розпізнавання білків з внутрішньою невпорядкованістю за допомогою методів машинного навчання. – Випускна кваліфікаційна робота магістра за спеціальністю 091 Біологія, ОП «Біоінформатика і структурна біологія».

Дослідження внутрішньо невпорядкованих білків та розробка алгоритмів передбачення структури таких білків є важливим для розуміння їх біологічних функцій, а також для вивчення багатьох захворювань. В даній роботі на основі існуючих алгоритмів-предикторів за допомогою статистичних методів побудовано модель метапредиктора з підвищеною якістю прогнозування окремих амінокислот білкової послідовності. Використовуючи результати метапредиктора, досліджено можливості застосування сучасних засобів автоматизованого машинного навчання для побудови моделі класифікації повних білкових послідовностей та отримано модель згорткової нейронної мережі з точністю на рівні 93%.

Исследование внутренне-неупорядоченных белков и разработка алгоритмов прогнозирования структуры таких белков важны для понимания их биологических функций, а также для изучения многих заболеваний. В данной работе на основе существующих алгоритмов-предикторов с использованием статистических методов построена метапредикторная модель с повышенным показателем прогнозирования отдельных аминокислот белковой последовательности. Используя результаты метапредиктора, исследованы возможности применения современных автоматизированных средств машинного обучения для построения модели классификации полных белковых последовательностей и получена модель сверточной нейронной сети с точностью 93%.

Understanding the biological functions of intrinsically disordered proteins and the development of algorithms for predicting their structure is essential for the study of numerous diseases. In this paper on the basis of existing predictor algorithms and statistical methods, a metapredictor model with improved quality of single amino acid prediction is created. Leveraging the results of the metapredictor, the possibility of using modern automated machine learning frameworks to construct a model for classifying entire protein sequences is evaluated, and a 93% accurate convolutional neural network model is developed.

Ключові слова: внутрішньо невпорядковані білки; предиктори IDP; метастратегія; машинне навчання.

ЗМІСТ

ВСТУП

Багато білків in vitro не мають стабільної фіксованої тривимірної структури або мають у своєму складі невпорядковані регіони. Вивчення білків з внутрішньою невпорядкованістю є важливим для інтерпретації та розуміння їх біологічних функцій, а також для вивчення багатьох захворювань.

Аналіз первинної структури великої кількості білків показав, що амінокислотні послідовності, не здатні згортатись до глобулярних структур, розповсюджені в природі дуже широко [1; 2; 3; 4]

За допомогою використання методів біоінформатики було встановлено, що багато білків, які мають відношення до таких патологій, як рак [5], діабет [6], серцево-судинні захворювання [7], нейродегенеративні захворювання [8] та інші, відносяться до класу повністю або частково невпорядкованих білків.

Завдяки наявності баз даних амінокислотних послідовностей та баз даних тривимірних структур білків з'явилась можливість використовувати методи аналізу даних та машинного навчання для розробки алгоритмів оцінки та передбачення білків з внутрішньою невпорядкованістю за їх амінокислотною послідовністю.

Не зважаючи на те, що на даний час розроблено більш ніж 100 алгоритмів-предикторів неструктурованих білків, багато з них не мають достатньо високої точності.

З огляду на це, підвищення якості передбачення білків з внутрішньою невпорядкованістю є надзвичайно актуальним завданням.

Мета даної роботи – дослідити можливості використання сучасних підходів та програмних засобів машинного навчання для підвищення якості передбачення білків з внутрішньою невпорядкованістю.

Відповідно до мети поставлено такі завдання:

1. Проаналізувати результати існуючих алгоритмів-предикторів та використати методи статистичної обробки для їх поєднання у метапредиктор з підвищеною якістю прогнозування приналежності окремих амінокислот білкової послідовності до класу невпорядкованих.
2. Дослідити можливості застосування сучасних засобів автоматизованого машинного навчання для побудови моделі класифікації повних білкових послідовностей.

РОЗДІЛ 1. ОГЛЯД ЛІТЕРАТУРИ

1.1. Просторова структура білка

Окрім первинної структури (послідовності амінокислот поліпептиду), для функціонування білка у клітині надзвичайно важливою є його просторова структура (конформація), яка формується в процесі згортання білків – фолдинга (англ. folding). Ця тривимірна конформація утримується завдяки взаємодії структур нижчих рівнів (рисунок 1.1). Така просторова структура білка за нормальних природних умов називається його нативним станом (або третинною структурою).

Згортання білка – це процес згортання синтезованого поліпептидного ланцюга в правильну просторову структуру. При цьому відбувається сходження віддалених залишків амінокислотного поліпептидного ланцюга, що призводить до утворення нативної структури. Ця структура має унікальну біологічну активність. Саме тому згортання є важливим етапом перетворення генетичної інформації в механізми функціонування клітин.

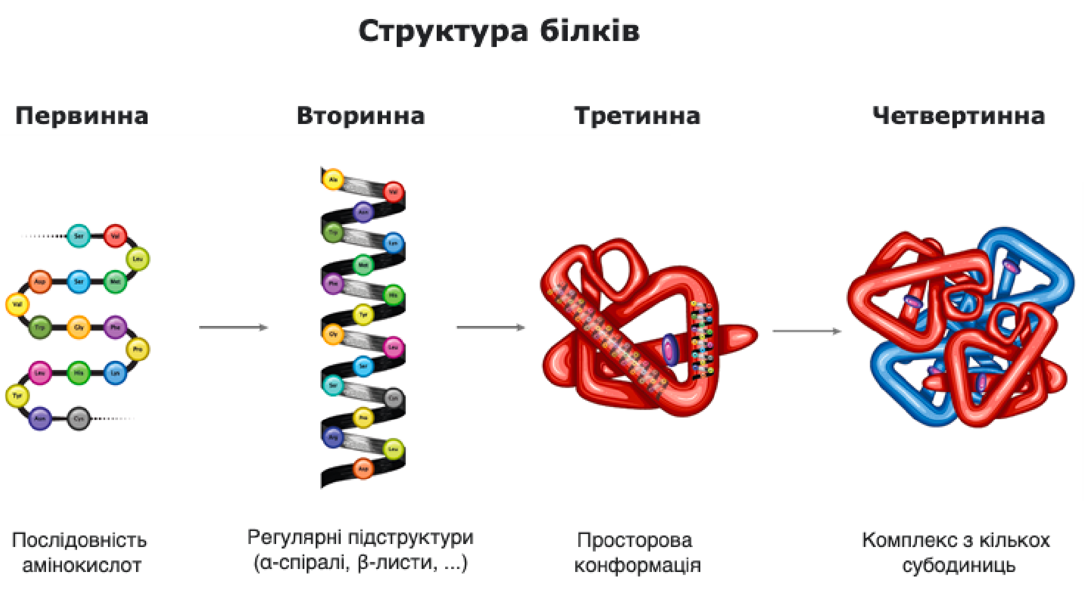

Рисунок 1.1 – Рівні організації білкових структур.

Для коректної роботи білків їх коректна просторова конформація є дуже важливою. При помилках згортання, як правило, утворюється неактивний білок з різними небажаними властивостями. Вважається, що деякі захворювання є наслідком накопичення неправильно згорнутих білків в клітині.

Міжмолекулярні взаємодії невпорядкованих ділянок нативних білків і білків в денатурованих частково складених проміжних станах можуть призводити до утворення аморфних асоціацій агрегатів, амілоїдів та амілоїдних фібрил.

1.2. Білки з внутрішньою невпорядкованістю

До недавнього часу в літературі переважала думка про те, що унікальна третинна структура білка необхідна для його функціонування.

Однак в останні роки було встановлено, що багато білків в клітині не володіють цією структурою в ізольованому стані, хоча і мають чітку біологічну функцію в фізіологічних умовах.

Такі білки отримали назву білків з природною або внутрішньою невпорядкованістю (IDP, intrinsically disordered proteins). Частка невпорядкованих регіонів (IDPR, intrinsically disordered regions) в таких білках може бути різною, починаючи від послідовності декількох амінокислот і закінчуючи повністю невпорядкованою послідовністю десятків, а іноді і сотень амінокислот.

Основна відмінність цих білків від структурованих (глобулярних) білків полягає в тому, що вони не мають унікальної третинної структури в ізольованому вигляді, а набувають її після взаємодії з партнерами. Їх конформація в комплексі визначається партнером взаємодії, а не тільки власною амінокислотною послідовністю, як це характерно для структурованих (глобулярних) білків [9-18].

Ознаками невпорядкування нативної структури білка можуть бути неможливість отримання кристалу білка, відсутність вираженої структури спектрів кругового дихроїзму в ближніх і (або) далеких УФ-областях спектру, великі гідродинамічні розміри макромолекули та ін. Основною проблемою при цьому є неможливість визначення координат атомів частини амінокислот поліпептидного ланцюга за допомогою рентгеноструктурного аналізу.

1.3. Існуючі методи передбачення білків з внутрішньою невпорядкованістю

Близько 70% структур банку даних Protein Data Bank (PDB) мають деякі невпорядковані залишки, а також близько 25% мають невпорядковані регіони (IDRs) більше 10 залишків у довжину [13]. Системні дослідження білків з внутрішньою невпорядкованістю (IDPs) виявили не тільки значну поширеність IDRs, але й їх зв'язок з різноманітними захворюваннями, включаючи рак, нейродегенеративні та серцево-судинні захворювання, амілоїдоз, діабет тощо [19].

Механізм згортання білка до кінця не вивчений. Експериментальне визначення тривимірної структури білка часто буває дуже складним та дорогим, в той час як амінокислотна послідовність білка зазвичай відома. Тому вчені намагаються використовувати інші підходи для того, щоб передбачити просторову структуру білка з його амінокислотної послідовності.

Розроблено багато обчислювальних методів, які широко використовуються для вивчення IDP та IDR, оскільки вони є більш ефективними та економічно вигідними у порівнянні з традиційними експериментальними методами [20].

На сьогоднішній день винайдено понад 100 алгоритмів-предикторів порушення структури білків [21], частина з яких втратила свою актуальність або не доступна для використання, а частина має недоліки внаслідок обмеженої інформації про характеристики білків в існуючих банках даних [22].

Опубліковані численні роботи, які дають історичний огляд, класифікують та описують предиктори порушення структури білків, а в деяких випадках порівнюють їх прогнозні показники [23-31].

Більш того, було проведено близько десяти масштабних порівняльних досліджень для оцінки прогнозних показників предикторів внутрішнього розладу [32-41]. Ці дослідження включають в себе декілька загальних відкритих конкурсів, де предиктори оцінювались на сліпих тестових наборах даних (тобто наборах даних, які не були доступні авторам предикторів) незалежними оцінювачами, які не беруть участі в конкурсах.

Серед таких конкурсів – «Критична оцінка прогнозування структури» (Critical Assessment of Structure Prediction, CASP) між CASP5 до CASP10, та «Критична оцінка внутрішнього білкового розладу» (Critical Assessment of Intrinsic Protein Disorder, CAID).

Найбільш актуальні результати CAID у 2021 році включають оцінку 32 алгоритмів-предикторів.

1.4. Стратегія покращення результатів класифікації

Однією з можливих стратегій вирішення проблеми недостатньої якості роботи існуючих алгоритмів-предикторів є використання метастратегії – поєднання результатів існуючих індивідуальних предикторів для покращення остаточного результату прогнозування [52]. Метастратегії були використані в декількох областях, таких як розпізнавання протеїнових укладок, прогнозування вторинної структури білка, взаємодії білка, субклітинної локації білка, посттрансляційних модифікацій, прогнозування промотера та в багатьох інших [53; 54].

Одним з завдань даного дослідження є розробка нової метастратегії за допомогою інтеграції окремих предикторів, які мають найвищу якість прогнозування.

Спочатку було відібрано найпопулярніші предиктори з тих, які широко використовуються в наукових дослідженнях і надають доступ до актуальних програмних пакетів або веб-серверів. Далі за допомогою регресійного аналізу були обрані найбільш значущі з предикторів. Предиктори, які мають найкращі показники, було інтегровано у якості предиктора метастратегії. Як наслідок, за допомогою машинного навчання було побудовано суттєво більш якісну модель класифікації та прогнозування ступеня невпорядкованості амінокислотних залишків білка.

1.5. Використання методів машинного навчання для аналізу білків

Машинне навчання (англ. Machine learning, ML) – клас методів штучного інтелекту, характерною рисою якого є не пряме вирішення проблеми, а «навчання» (поступове покращення продуктивності у деякій задачі) за допомогою застосування рішень багатьох подібних проблем. Для побудови таких методів застосовуються інструменти математичної статистики, чисельні методи, математичний аналіз, методи оптимізації, теорія ймовірностей.

Еволюціювавши з досліджень теорії обчислювального навчання та розпізнавання образів, машинне навчання зараз застосовують у ряді задач, для яких розробка явних алгоритмів з бажаною якістю є дуже складним або неможливим завданням.

Багато з відомих методів машинного навчання використовуються і для аналізу біологічних послідовностей.

Lee et al. [42] запропонував альтернативний алгоритм дерева рішень для оцінки надійності взаємодії білка. Geng et al. [43] запропонував метод з використанням Naive Bayes для прогнозування ділянок взаємодії білок-білок. Qi et al. [44] представляють новий метод з використанням алгоритму випадкових лісів (Random Forest) для обчислення ступеня подібності білків та класифікації пар білків як взаємодіючих чи ні. Ramkumar et al. [45] запропонував багатошаровий персептронний підхід для прогнозування білкових вторинних структур з використанням різного набору вхідних функцій та параметрів в розподіленому обчислювальному середовищі.

Глибоке навчання (англ. Deep learning, DL) – це сукупність підходів, що застосовуються в машинному навчанні для моделювання високорівневих абстракцій даних за допомогою архітектур, в основі яких велика кількість нелінійних перетворень. Різні архітектури глибинного навчання, такі як згорткові глибинні нейронні мережі, знайшли застосування в таких областях, як автоматичне розпізнавання та обробка природної мови, комп'ютерне бачення та біоінформатика, де вони демонструють відмінні результати в різноманітних задачах.

Згорткова нейронна мережа (англ. Convolutional neural network, CNN) – спеціальна архітектура штучних нейронних мереж, запропонована Яном Лекуном у 1988 році, націлена на ефективне розпізнавання образів та успішно зарекомендувала себе в завданнях розпізнавання та класифікації зображень [46].

Застосування згорткових нейронних мереж [49] до геномних послідовностей у 2015 році сигналізувало про появу епохи глибокого навчання в комп'ютерній біології. Останнім часом вчені також використовують CNN для прогнозування білкових структур, наприклад для ДНК-білкового зв'язування [47]. Haoyang Zeng та співавтори [48] визначили найкращі архітектури CNN, змінюючи параметри, глибину та розміри фільтрів. Два методи, DeepBind [50] і DeepSea [51], успішно застосували глибоке навчання для моделювання специфіки зв'язування білка та продемонстрували ефективність, що перевершує кращі існуючі традиційні методи навчання.

1.6. Засоби автоматизації машинного навчання

Зі збільшенням кількості завдань і областей застосування машинного навчання виникають і нові виклики. Традиційна розробка моделей машинного навчання займає багато ресурсів, вимагає значного обсягу знань в предметній області і часу для створення та порівняння десятків моделей.

В останні роки відбуваються фундаментальні зміни в підходах до машинного навчання і науки про дані – автоматизація самого процесу машинного навчання. Ця концепція виникла головним чином через те, що застосування традиційних методів машинного навчання до реальних рішень є трудомістким і часто складним навіть для експертів. Для цього потрібні знання, навички, досвід і наявність професіоналів або експертів з різних дисциплін. Концепція автоматизованого машинного навчання робить науку про дані більш доступною для всіх, використовуючи кращі практики машинного навчання від провідних вчених даних з усього світу.

Автоматизоване машинне навчання (англ. Automated machine learning, AutoML) – це процес автоматизації трудомістких і повторюваних завдань розробки моделей машинного навчання. Він може бути використаний для створення моделей машинного навчання з високою масштабованістю, ефективністю та продуктивністю.

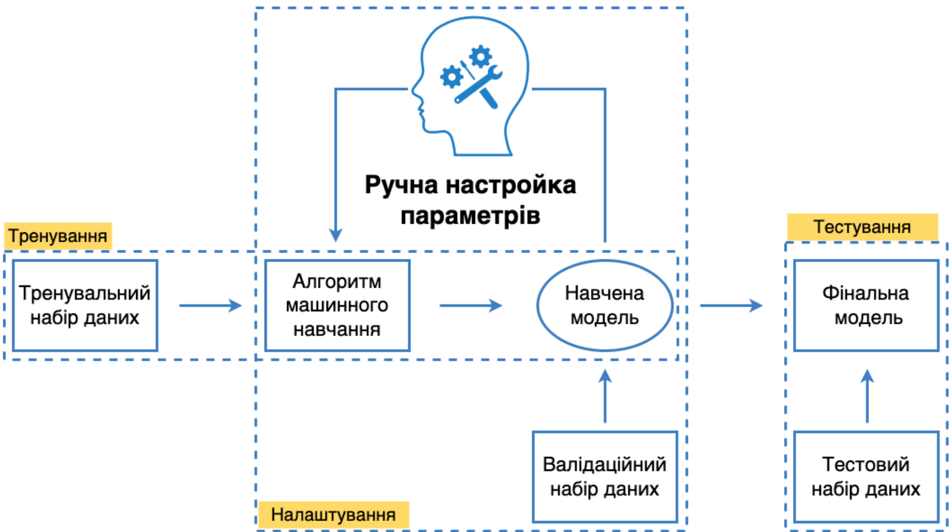

Рисунок 1.2 – Типовий процес машинного навчання.

У типовому застосуванні машинного навчання, спеціалісти мають набір вхідних даних для тренування моделі. Ці дані можуть не бути у вигляді, до якого можливо безпосередньо застосувати потрібні алгоритми. Для приведення цих даних у форму, придатну для машинного навчання, може бути необхідно використати відповідні методи попередньої обробки даних, конструювання, виділення та обрання ознак. Після цих кроків експерти мусять власноруч обрати алгоритм, архітектуру моделі та виконати оптимізацію гіперпараметрів, щоб максимізувати передбачувальну продуктивність своєї моделі (рисунок 1.2). Кожен із цих кроків може виявлятися складним, спричиняючи значні перешкоди для використання машинного навчання. AutoML різко спрощує перелічені кроки для неекспертів, автоматизуючи всі ці процеси.

Рисунок 1.3 – Процес автоматизованого машинного навчання.

Рівень автоматизації в AutoML дозволяє використовувати методи та моделі машинного навчання, не вимагаючи від користувача високого рівня експертності в машинному навчанні. Автоматизація кожного кроку процесу машинного навчання дозволяє швидше створювати рішення та моделі, які часто можуть бути простішими та перевершувати за ефективністю рішення, розроблені власноруч (рисунок 1.3).

Дослідники в галузях медицини, нейробіології, геноміки та інших тепер можуть запропонувати нові рішення специфічних для домену проблем, таких як сегментація медичних зображень, геномні дослідження та аналіз біологічних даних, не проходячи довгий шлях навчання ML та програмуванню.

В даній роботі було досліджено можливості застосування фреймворку автоматизованого машинного навчання AutoKeras для побудови моделі класифікації повних білкових послідовностей, а результати автоматично побудованих моделей співставлені з результатами роботи деяких моделей, сконструйованих вручну.

РОЗДІЛ 2. МЕТОДИ ТА РЕЗУЛЬТАТИ ДОСЛІДЖЕННЯ

2.1. Об’єкт, матеріали та методи дослідження

2.1.1. Об’єкт дослідження

Об'єктом даного дослідження виступають внутрішньо невпорядковані білки, інформацію щодо яких можна отримати з відкритих джерел.

В даний час існує кілька банків даних, що містять інформацію про амінокислотні послідовності білків з внутрішньою невпорядкованістю, отримані експериментальним шляхом. Серед них DisProt, PDB, MobiDB, IDEAL, DIBS та MFIB. Однак ці ресурси охоплюють лише незначну частину внутрішньо невпорядкованих білків, а найбільші бази даних DisProt та PDB наразі включають близько 2000 та 25000 найменувань відповідно.

В порівнянні з більш ніж 225 мільйонами білкових послідовностей, доступних у UniProt станом на 2021 рік, для всебічної ідентифікації та анотації білків з внутрішньою невпорядкованістю потрібно пройти ще довгий шлях.

Обчислювальні методи, які точно передбачають ступінь внутрішньої невпорядкованості, можуть бути використані для полегшення зусиль, спрямованих на закриття цього величезного і зростаючого розриву в знаннях. Існуючі алгоритми-предиктори вже зробили свій внесок та в значній мірі вплинули на прискорення досліджень внутрішньо невпорядкованих білків. Вони також використовуються у багатьох сферах, включаючи медицину, раціональний дизайн ліків та структурну геноміку.

2.1.2. Матеріали

2.1.2.1 Набори даних

У цьому дослідженні використано три незалежні набори даних з бази DisProt (http://www.disprot.org) та RCSB Protein Data Bank (PDB) [55].

Перший навчальний набір був підготовлений організаторами в рамках змагання Critical Assessment of Intrinsic Protein Disorder (CAID) на основі банку даних білкових розладів DisProt і включає послідовності, анотовані протягом червня – листопада 2018 року, версія «2018_11 (CAID)» [35].

Другий навчальний набір отримано з бази даних RCSB Protein Data Bank (PDB) від 28 лютого 2017 року шляхом вибору тільки одноланцюгових послідовностей при наявності 15 або більше моделей.

Набір для незалежного тестування складається з даних рентгенівської кристалографії банку PDB, який був оновлений 2 березня 2017 року.

2.1.2.2 Індивідуальні предиктори

Розробка предикторів внутрішньої невпорядкованості є давньою дослідницькою проблемою. За результатами огляду 2021 року відомо більше 100 індивідуальних предикторів, які були створені протягом останніх 40 років [21]. Сучасні дослідження вказують на довгу історію в області прогнозування IDP, надаючи безцінну інформацію про архітектуру методів, їх доступність, тенденції в їх розробці та порівняльну оцінку їх прогнозної ефективності.

Архітектури та методи, що використовуються для розробки предикторів внутрішньої невпорядкованості, зазвичай поділяються на три категорії:

1. Функції оцінки послідовності;
2. Моделі машинного навчання;
3. Метапредиктори.

Перша категорія використовує адитивні та/або зважені функції, деякі з яких обґрунтовані фізичними принципами, що регулюють згортання білка, для обробки вхідної амінокислотної послідовності білка разом із структурною та еволюційною інформацією. Предикторі, які належать до цієї категорії, включають FoldIndex та IUPred. Предиктори другої категорії застосовують моделі, отримані за допомогою різних алгоритмів машинного навчання, таких як метод опорних векторів (Support vector machines, SVM), регресія, умовні випадкові поля (Conditional random fields, CRFs), мережі радіально базисних функцій (Radial basis function network, RBF) та неглибокі нейронні мережі (shallow neural networks). Прикладом популярних предикторів з категорії машинного навчання є DisEMBL, DISOPRED, PONDR та PRDOS. Метапредиктори використовують кілька результатів прогнозу як вхідні дані для виконання повторного прогнозування. Прикладами метапредикторів є MFDp, Cspritz, MobiDB-Lite.

Однак останнім часом відзначається швидке поширення нової підродини методів машинного навчання, яка спирається саме на глибокі нейронні мережі після того, як в 2013 році був випущений перший метод на основі CNN. Предиктори на основі CNN відрізняються від неглибоких нейронних мереж, які зазвичай використовувались для створення предикторів розладів на початку 2000-х років, використанням декількох прихованих шарів та більш складних типів нейронів та їх з'єднань. Перехід до моделей глибокого навчання обгрунтований їх високим рівнем прогнозної продуктивності відносно інших методів. Як можна бачити з результатів нещодавно проведеного змагання CAID, методи, що демонструють найкращі показники, розроблені саме за допомогою підходів глибокого навчання. Серед них flDPnn, SPOT-Disorder2, RawMSA та AUCPred.

2.1.2.3 AutoML фреймворк

Доступні інструменти AutoML можна розділити на кілька категорій в залежності від їх основного фокусу:

1. Інструменти для автоматизованого конструювання ознак
2. Інструменти для автоматизованої настройки гіперпараметрів, вибору моделі та генерування пайплайнів (pipelines)
3. Інструменти для автоматизованого глибокого навчання

Переважна більшість існуючих проектів AutoML фокусується на налаштуванні гіперпараметрів або генерації пайплайнів ML.

В останні роки більшість зусиль з розробки зосереджується на автоматизованому глибокому навчанні. Список існуючих проектів включає:

- AutoKeras (https://autokeras.com)
- Auto-Gluon (https://auto.gluon.ai)
- TPOT (http://epistasislab.github.io/tpot/)
- Microsoft NNI (https://nni.readthedocs.io)
- H2O AutoML toolkit (https://h2o.ai/platform/h2o-automl/)
- Ludwig (https://github.com/ludwig-ai/ludwig)
- LightAutoML (https://github.com/AILab-MLTools/LightAutoML)

Окрім проектів з відкритим кодом (opensource), багато компаній, особливо тих, що надають хмарні послуги, також розробляють комерційні платформи AutoML. Серед таких:

- Google Cloud AutoML (https://cloud.google.com/automl)
- Amazon SageMaker Autopilot (https://aws.amazon.com/sagemaker/autopilot/)
- Microsoft Azure AutoML (https://azure.microsoft.com/en-us/services/machine-learning/automatedml/)
- IBM Watson Studio AutoAI (https://www.ibm.com/cloud/watson-studio/autoai)
- H2O.ai (https://h2o.ai/platform/ai-cloud/make/h2o-driverless-ai/)

Для вибору інструменту AutoML в рамках даної роботи було застосовано такі критерії:

- проект з відкритим кодом (open source)
- проект широко використовується, має розвинуте ком'юніті та гарну документацію
- проект дозволяє будувати моделі глибокого навчання

- проект дозволяє гнучке налаштування гіперпараметрів та властивостей архітектури для автоматизованої побудови оптимальної моделі

Серед наявних в даних час інструментів було обрано AutoKeras – бібліотеку Python, орієнтовану на автоматизовану генерацію моделей глибокого навчання. Проект AutoKeras побудований на основі TensorFlow (https://tensorflow.org), TensorFlow Keras API (https://keras.io) та бібліотеки KerasTuner (https://keras.io/keras_tuner).

TensorFlow – це розвинута платформа машинного навчання з відкритим кодом. Вона включає гнучку екосистему інструментів і бібліотек, які дослідники та розробники можуть використовувати для створення і розгортання ML-додатків. TensorFlow реалізує повний набір математичних операцій для роботи на різних апаратних засобах, включаючи процесори, графічні процесори та блоки обробки тензорів (Tensor processing unit, TPU) для глибокого навчання моделі. Навчання може бути масштабовано до декількох графічних процесорів на декількох машинах, а отримана модель може бути розгорнута в різних середовищах, таких як веб-сторінки та вбудовані системи.

Keras – це бібліотека Python, яка надає більш простий набір API для побудови та навчання моделі, інкапсулюючи функціональність TensorFlow. Це значно зменшує зусилля, необхідні для створення алгоритмів глибокого навчання. Keras виник як окремий пакет Python, але був інтегрований у пакет TensorFlow як API високого рівня, що полегшує налаштування, масштабування та розгортання моделей глибокого навчання.

AutoKeras позиціонується як бібліотека найвищого рівня з усіх бібліотек в екосистемі Keras. Вона пропонує найвищий рівень автоматизації для створення рішення глибокого навчання для обраної цільової задачі ML, наприклад для класифікації зображень.

2.1.3. Методи

2.1.3.1 Підготовка даних

Перший набір даних є референсним набором з банку даних білкових розладів DisProt і включає 652 послідовності з невпорядкованими амінокислотними залишками. Всі амінокислотні залишки, що не мали анотації DisProt як неструктуровані, вважались такими, що відносяться до структурованих регіонів. Кількість невпорядкованих амінокислот в цьому наборі даних склала 54820 при загальній кількості в 337908 залишків.

В другому навчальному наборі структури, що мають ліганди або дисульфідні зв'язки, були вилучені з набору даних. Послідовності, аналогічні послідовностями з першого набору, було видалено. Далі послідовності набіру були згруповані за допомогою програми CD-HIT (http://weizhong-lab.ucsd.edu/cd-hit) з параметром ідентичності послідовностей 30%. Причина використання рівня відсічки у 30% полягає в тому, що зазвичай послідовності, що подібні менш ніж на 30%, вважаються функціонально та еволюційно не пов'язаними. Застосовуючи розподіл значення RMSD кожного залишку, порогове значення 5 було обрано як рівень відсічки для визначення того, чи є цей залишок невпорядкованим або структурованим. Як наслідок, отримано 3155 ланцюгів, а кількість невпорядкованих залишків в порівнянні з усіма залишками в цьому наборі даних становить 36179 до 326433.

З набору для незалежного тестування було видалено послідовності, спільні з першим та другим тренувальним набором. Після видалення набір даних склав 10505 ланцюгів та 2578418 залишків амінокислот. Кількість невпорядкованих амінокислот в цьому наборі даних склала 142750.

У якості попередньої обробки даних для кожної амінокислоти отримувався прогноз за допомогою обраних предикторів. Кожне прогнозоване значення розглядалось як незалежна змінна для цього екземпляра амінокислоти. Кожну амінокислоту було позначено як «D» або «O», де «D» означає, що ця амінокислота невпорядкована або знаходиться в невпорядкованому регіоні. «O» означає, що ця амінокислота впорядкована та знаходиться в добре структурованому регіоні.

2.1.3.2 Вибір предикторів

Шляхом пошуку в PubMed, Web of Science та аналізу наукових статей було сформовано набір з 42 предикторів. Першим кроком відбору окремих предикторів був огляд наявної літератури та перевірка працездатності веб-сервера.

Далі для відбору застосовувались такі критерії:

1) Предиктор широко використовується і цитується. Обирались тільки предиктори з оригінальними науковими роботами, що цитуються більше 10 разів на рік.

2) Предиктор має актуальну версію та добре підтримуваний працездатний веб-ресурс.

В результаті з початкового набору було покроково відібрано 13 предикторів та отримано набір для побудови логістичної регресійної моделі, який включає як ті предиктори, що продемонстрували найкращі показники за даними звіту про змагання CAID та засновані на підходах глибокого навчання, так і відомих представників інших категорій.

Серед відібраних предикторів:

- flDPnn (http://biomine.cs.vcu.edu/servers/flDPnn/)
- RawMSA (https://bitbucket.org/clami66/rawmsa)
- Espritz (http://old.protein.bio.unipd.it/espritz)
- SPOT-Disorder2 (https://sparks-lab.org/server/spot-disorder2/)
- DisoMine (https://www.bio2byte.be/b2btools/disomine/)

- IUPred (https://iupred2a.elte.hu)
- DisEMBL (http://dis.embl.de)
- RONN (https://www.bioinformatics.nl/~berndb/ronn.html)
- PONDR-VXLT (http://www.pondr.com)
- PONDR-VSL2 (http://www.pondr.com)
- PONDR-FIT (http://original.disprot.org/pondr-fit.php)
- Globplot (http://globplot.embl.de)
- Dispro (http://scratch.proteomics.ics.uci.edu)

PONDR-VSL2, DisEMBL та Globplot засновані на штучній нейронній мережі [57; 58]. PONDR-VXLT базується на методі опорних векторів [57]. Dispro використовує комбінацію нейронних мереж та байєсівських методів [57]. IUPred, Espritz і RONN розроблені на базі як фізичних методів, так і нейронних мереж [58; 59]. AUCpreD базується на CNN з помірною глибиною [63]. flDPnn заснований на архітектурі нейронної мережі з прямим зв'язком (Feedforward neural network, FFNN) помірної глибини [60]. SPOT-Disorder2 та RawMSA [61; 62] використовують дуже глибоку архітектуру гібридних двоспрямованих рекурентних нейронних мереж (Bidirectional recurrent neural networks, BRNN) та CNN.

2.2. Результати досліджень

2.2.1. Логістична модель регресії для вибору окремих предикторів

Насамперед необхідно було визначити найбільш значущі предиктори зі списку обраних. Спочатку було перевірено ефективність кожного окремого предиктора шляхом оцінки його точності та частоти істинно позитивних результатів.

Було визначено, що індивідуальні предиктори мають діапазон частоти істинно позитивних результатів (TP Rate) від 39.3% до 78.6%. Додатково було перевірено точність кожного предиктора на підготовлених тестових даних. Точність обраних предикторів становить близько 63.5% – 80.1%, як показано в таблиці 2.1, що є недостатньо високим показником довіри.

Предиктор	**Частота TP**	**Частота FP**	**ROC крива**	**Точність**	**R^2**	**p-значення**
flDPnn	0.646	0.227	0.811	80.1%	0.6071	<0.0001
RawMSA	0.380	0.159	0.754	78.2%	0.6337	<0.0001
IUPred	0.451	0.107	0.732	76.7%	0.6891	<0.0001
SPOT-Disorder2	0.732	0.387	0.688	77.1%	0.7444	<0.0001
Espritz	0.605	0.420	0.802	78.4%	0.7976	<0.0001
DisEMBL	0.393	0.366	0.679	72.6%	0.8342	0.005
VSL2	0.606	0.320	0.701	65.3%	0.8583	0.008
VLXT	0.786	0.409	0.733	64.7%	0.8712	0.012
DisoMine	0.639	0.377	0.712	71.0%	0.8676	0.432
RONN	0.499	0.311	0.735	71.4%	0.8523	0.552
PONDRFIT	0.549	0.121	0.743	73.7%	0.8398	0.601
Dispro	0.644	0.288	0.632	68.2%	0.7921	0.658
Globplot	0.752	0.225	0.602	63.5%	0.7409	0.697

Таблиця 2.1 – Список індивідуальних предикторів.

Щоб обрати найбільш значущі предиктори було використано метод прямого покрокового логістичного вибору (logistic forward selection). Метод передбачає наявність початкової моделі без змінних, тестування додавання кожної змінної за допомогою критерію відповідності вибраної моделі, додавання змінної (якщо є), включення якої дає найбільш статистично значне поліпшення, та повторення цього процесу до тих пір, поки жодна з доданих змінних не зможе покращити модель з достатньою ступінню статистичної значимості.

$$\log_e\left(\frac{p(D)}{p(O)}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ \dots\ + \beta_{13} x_{13}$$

Формула 2.1 – Модель логістичної регресії.

В цій моделі *p(D)* – це ймовірність того, що амінокислотний залишок буде класифіковано як невпорядкований; *p(O)* становить *1−p(D)*, що є ймовірністю класифікації залишку як упорядкованого. Значення β_0 – β_{13} є модельними оцінками, x_1 – x_{13} є прогнозованими значеннями для кожного амінокислотного залишку. Для того, щоб отримати найбільш значущу модель, використано найбільш конвергентний набір даних DisProt, який має 408691 екземпляри амінокислотних залишків.

Покрокова послідовність показана як послідовність рядків у таблиці 1. IUPred є найбільш значущим предиктором для моделі і обирається першим, потім обирається Disemble, Espritz і так далі. Внесок кожного предиктора в рівняння моделі (формула 2.1) можна обчислити за допомогою коефіцієнта детермінації (R^2).

Було виявлено, що після того, як логістична модель включає перші 8 предикторів, коефіцієнт детермінації R^2 досягає пікового значення 0.8712. Якщо додати інші 3 предиктори до моделі, R^2 буде поступово зменшуватись, як показано в таблиці 1. Це свідчить про те, що останні три предиктори мають негативний вплив на перші 8 доданих. Тому було прийнято рішення використовувати тільки перші 8 предикторів у побудові моделі машинного навчання для підвищення точності.

2.2.2. Побудова класифікатора окремих амінокислотних залишків

Як вказано у розділі 2.1.3.1, на етапі підготовки даних кожний амінокислотний залишок було марковано як «D» або «O». Тренувальний набір даних склав 408691 елементів з 8 атрибутами (значення прогнозування амінокислотного залишку кожним з 8 обраних індивідуальних предикторів).

На підготовлених даних за допомогою програмного забезпечення Weka (https://www.cs.waikato.ac.nz/ml/weka) було проведено перехресне затверджування (крос-валідація, cross-validation) – оцінку аналітичної моделі та її поведінки на незалежних даних. При перехресному затверджуванні наявні дані розбиваються на k частин (k-fold cross-validation). Потім модель навчається на k−1 фрагментах даних, а решта використовується для валідації (рисунок 2.1). Процедура повторюється k разів, в результаті кожна з k частин даних використовується для валідації. Результатом є оцінка ефективності обраної моделі з максимально рівномірним використанням наявних даних.

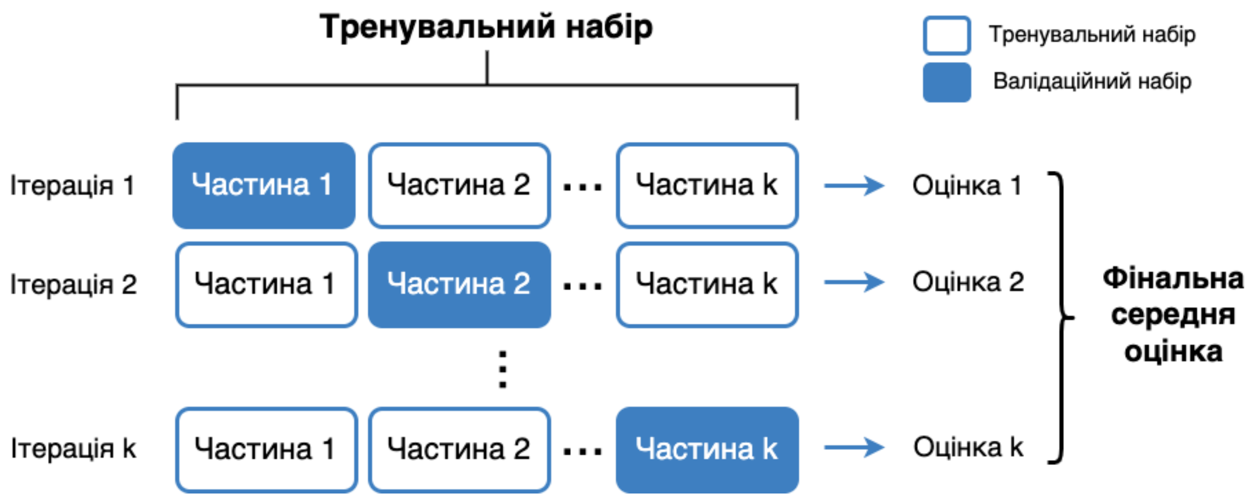

Рисунок 2.1 – Процедура перехресного затверджування.

На рисунку 2.2 зображено схему процесу перехресного затверджування (10-fold) з використанням 4 різних класифікаторів:

- Decision tree J48
- Naive Bayes
- Random Forest
- Multilayer Perceptron (Neural Network)

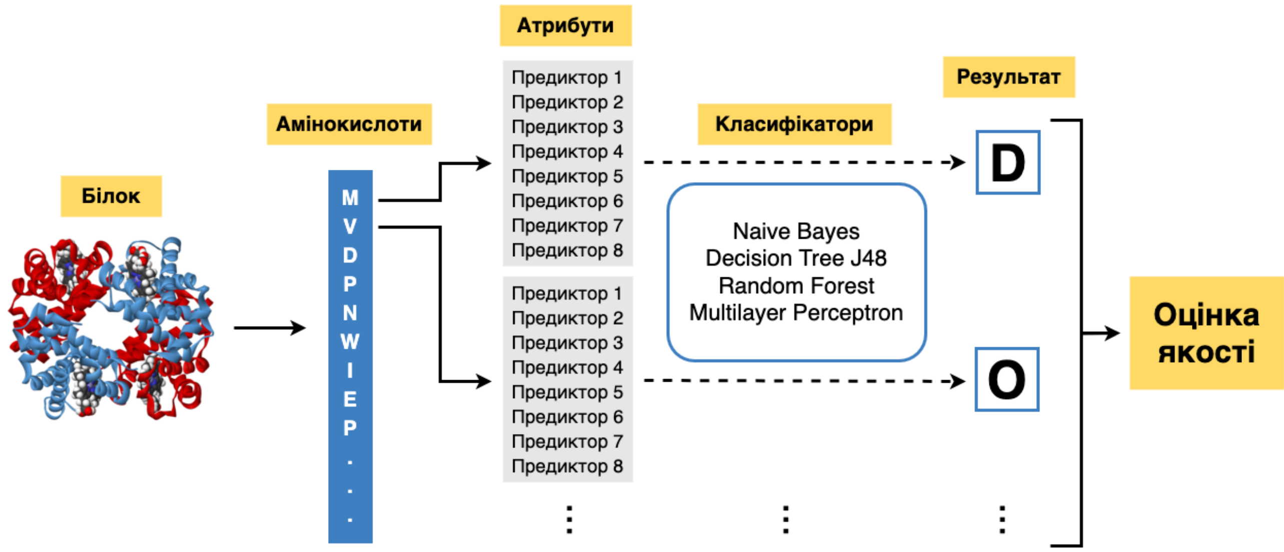

Рисунок 2.2 – Процес оцінки класифікаторів.

Матрицю невідповідності зображено в таблиці 2.2.

Класифікатор	Naive Bayes		Decision Tree J48		Random Forest		Multilayer Perceptron	
Результат	D	O	D	O	D	O	D	O
D	57760	33472	60724	32656	68942	19916	46154	47230
O	56136	261323	17764	297547	7263	312570	25789	289518

Таблиця 2.2 – Матриця невідповідності класифікаторів.

Точність J48 склала 86.15%, Naive Bayes – 77.11%, Random Forest – 92.21%, Multilayer Perceptron – 83.29%.

Окрім точності для порівняння були використані інші метрики, такі як частота хибно позитивних результатів, F-міра, коефіцієнт кореляції Меттьюза (MCC) та ROC-крива. Результати порівняння відображені в таблиці 2.3.

Класифікатор	Частота TP	Частота FP	Чутливість	Специфічність	F-міра	MCC	ROC	Точність
Naive Bayes	0.742	0.322	0.811	0.773	0.778	0.430	0.818	77.11%
Decision Tree J48	0.861	0.288	0.889	0.886	0.869	0.669	0.869	86.15%
Random Forest	0.928	0.179	0.945	0.944	0.932	0.805	0.967	92.21%
Multilayer Perceptron	0.802	0.465	0.810	0.832	0.812	0.451	0.845	83.29%

Таблиця 2.3 – Порівняння результатів класифікаторів.

Частоти хибно позитивних (FP Rate) та істинно позитивних (TP Rate) результатів широко використовуються в машинному навчанні в якості міри оцінки. Чим вище значення TP Rate – тим краще, чим нижче значення FP Rate – тим краще. Чутливість (Precision) вимірює частку істинно позитивних результатів, що є визначеними правильно. Специфічність (Recall) вимірює правильно визначену частку істинно негативних результатів. F-міра (F-Measure) – одна з мір точності тесту в статистичному аналізі бінарної класифікації. Коефіцієнт кореляції Меттьюза (Matthews correlation coefficient, MCC) використовується в машинному навчанні як міра якості бінарної класифікації, де значення, найбільш близьке до 1, означає більшу точність передбачення. ROC-крива (receiver operating characteristic, робоча характеристика приймача) – графік, що дозволяє оцінити якість бінарної класифікації, відображає співвідношення між часткою об'єктів від загальної кількості носіїв ознаки, правильно класифікованих, до загальної кількості об'єктів, що не несуть ознаки помилково класифікованих як такі, що мають ознаку. Також відома як крива похибок.

Як можна бачити, Random Forest має найвищі показники, а отже демонструє найкращу ступінь точності серед протестованих класифікаторів.

2.2.3. Побудова класифікатора повної амінокислотної послідовності

Якщо розглядати послідовність даних білка як зображення, можна використати згоркову нейронну мережу (CNN) для класифікації білкових даних – замість обробки піксельних даних зображення можливо обробляти послідовність характеристик білка. Оскільки біологічні дані можуть мати величезні об'єми, CNN є ефективним інструментом для їх аналізу.

Згорткові нейронні мережі є варіацією багатошарових перцептронів, розроблених для імітації функціонування зорової кори головного мозку. Основна ідея CNN полягає у екстрагуванні деяких локальних патернів зображення і поступове їх узагальнення у більш складні патерни шар за шаром. Архітектура CNN формується набором спеціалізованих шарів, які дозволяють відобразити вхідні дані на множину вихідних класів за допомогою диференційованих функцій.

Базова архітектура згорткової нейронної мережі (рисунок 2.3) складається з чотирьох основних типів шарів:

1. згортковий шар (convolution layer);
2. шар лінійного випрямлення (Rectified Linear Unit, ReLU);
3. шар субсемплінгу або пулінгу (pooling layer);
4. повноз'єднаний шар (fully connected layer, dense layer).

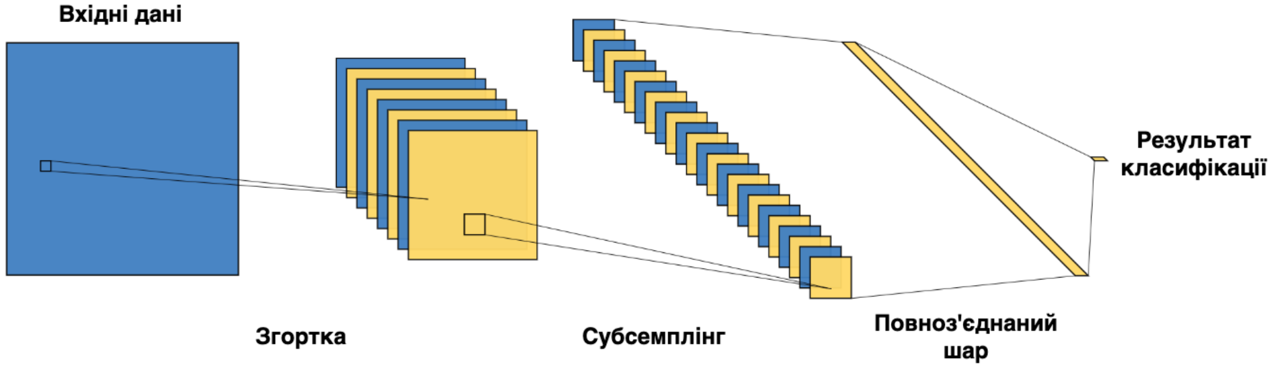

Рисунок 2.3 – Типова архітектура згорткової нейронної мережі.

Шар згортки – основа нейронної мережі, представлений набором тривимірних матриць, які називають фільтрами. За допомогою кожного фільтра відбувається операція згортки матриці вихідних даних. Матриця, отримана після застосування згорткового фільтра, ще називається картою ознак (feature map). Ця матриця відповідає деякому паттерну, знайденому у вхідних даних.

Шар лінійного випрямлення (шар активації) є додатковим шаром. Він застосовується після кожного згорткового шару і містить функцію активації — поелементну операцію, що застосовується до кожного елементу даних. Зазвичай в якості функції активації використовують ReLU, але існують і інші (наприклад, tanh або sigmoid).

Шар субсемплінгу призначений для зменшення розмірності кожної карти ознак зі збереженням найважливішої інформації про вхідні дані. Крім того, субсемплінг зменшує кількість параметрів нейронної мережі і, як наслідок, необхідну кількість обчислень. Таким чином, шар субсемплінгу служить для запобігання проблеми перенавчання (overfitting). Операції, що виконуються на цьому шарі, реалізуються за допомогою різних функцій: max, avg, sum та ін.

Повноз'єднаний шар – це традиційний шар багатошарових нейронних мереж, який використовує функцію активації softmax у вихідному шарі. Цей шар повністю зв'язаний, тобто кожен нейрон попереднього шару з'єднаний з кожним нейроном наступного. У нейронній мережі останній шар цього типу формує високорівневі ознаки вхідних даних. Ці ознаки використовуються для класифікації на основі навчального набору даних. Додавання повноз'єднаного шару зменшує обчислювальні витрати, необхідні для навчання мережі. Сума ймовірностей, які отримані в результаті обчислень в повноз'єднаному шарі, повинна дорівнювати 1.

Зазвичай до складу архітектури мережі також включають шар втрат (dropout layer). Шар втрат випадково змінює деякі з вхідних елементів на нуль під час кожної навчальної ітерації, маскуючи таким чином деякі з нейронів – під час кожного прямого проходу і зворотного поширення взаємодіють і оновлюються тільки немасковані нейрони. Це зменшує рівень кореляцій між нейронами і вносить додаткову випадковість в навчальний процес, що може допомогти мережевим шарам адаптуватися до різноманітних вхідних даних та покращити їх здатність до узагальнення. Шар втрат застосовується тільки в процесі навчання, щоб уникнути проблеми перенавчання мережі.

Для побудови класифікатора повної амінокислотної послідовності білка за допомогою CNN, спочатку необхідно було визначитися з кількістю класів для маркування кожної з послідовностей навчального набору відповідною міткою, яка буде використана в якості цільового значення (target) в процесі навчання моделей.

В якості міток було обрано 4 класи на основі співвідношення кількості неструктурованих амінокислот до довжини повної послідовності, відповідно до стандарту класифікації внутрішньо невпорядкованих білків [66]. Якщо білок має більше 80% неструктурованих амінокислотних залишків, він позначається як клас 1, білок з високим ступенем невпорядкованості. Від 50% до 80% – клас 2, білок середнього ступеня невпорядкованості. 25% до 50% – клас 3, білок з низьким ступенем невпорядкованості. При наявності 25% або меншої кількості неструктурованих залишків білок позначається як клас 4, впорядкований білок. Для підрахунку кількості неструктурованих амінокислотних залишків кожного білку було використано класифікатор на основі Random Forest, отриманий у розділі 2.2.2.

При використанні двомірної CNN для класифікації амінокислотної послідовності білка вхідні дані для нейронної мережі потрібно надавати у вигляді матриці. Для кожного білка його амінокислотну послідовність було перетворено на вектор – послідовність атрибутів, де кожна амінокислота представлена вісьмома прогнозними значеннями обраних індивідуальних предикторів. Так як довжина ланцюгів білків с тренувального набору даних коливається від 50 до більш ніж 1800 залишків, було обрано довжину вектора 14400, якої достатньо, щоб вмістити атрибути 1800 амінокислот. Якщо білок коротший за 1800 залишків, вектор доповнюється нульовими атрибутами до встановленої довжини 14400. Якщо білок довший за 1800 залишків, до вектору включаються тільки атрибути перших 1800 амінокислот. Після побудови вектора його можна представити у вигляді матриці розмірністю 120 × 120 (рисунок 2.4).

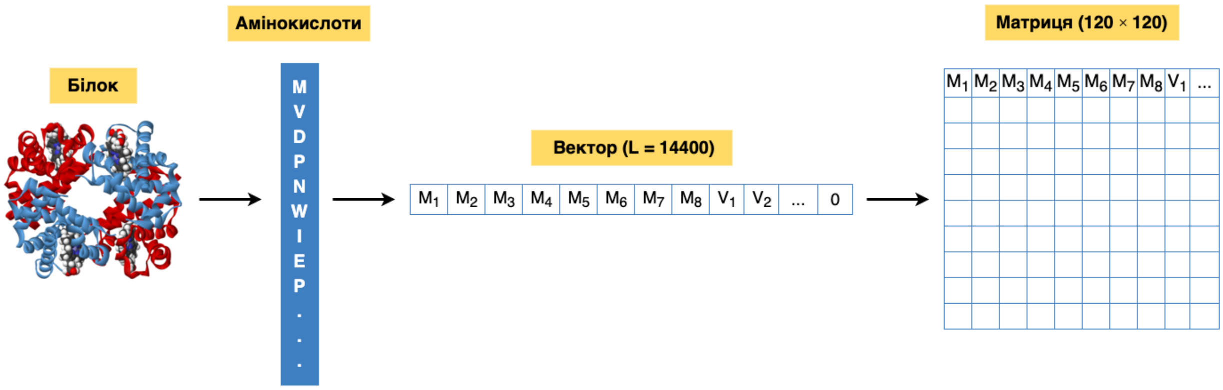

Рисунок 2.4 – Перетворення вхідних даних на матрицю.

Можливості бібліотеки AutoKeras дозволяють проектувати високорівневий дизайн мережі за допомогою згоркових блоків, кожен з яких послідовно включає низку згорткових шарів, шар субсемплінгу та шар втрат. Після згорткових блоків додається повноз'єднаний шар.

Пошуковий простір для автоматичного підбору оптимальних гіперпараметрів спроектованої мережі формується на основі наступних значень:

- кількість згорткових блоків
- кількість згорткових шарів в кожному блоці
- кількість фільтрів згорткового шару
- розмір ядра (kernel) фільтру
- дозвіл на застосування шару субсемплінгу
- дозвіл на застосування повноз'єднаного шару
- розмір повноз'єднаного шару

З огляду на наявні дослідження найкращих архітектур CNN [48; 64; 65], було обрано наступний діапазон значень для пошукового простору гіперпараметрів:

- кількість згорткових шарів – [1, 2]

- кількість фільтрів згорткового шару – [2, 4, 8, 16, 32, 64]
- розмір ядра фільтру – [3, 5, 7]
- розмір повноз'єднаного шару – [256, 512]

В шарі субсемплінгу виконувалось максимізаційне агрегування (max pooling) – використовувалось максимальне значення з кожного кластеру нейронів попереднього шару при розмірі кластеру 2 × 2. Значення 0.5 використовувалось в якості коефіцієнта у шарі втрат.

Так як при такій конфігурації пошукового простору кількість можливих варіантів CNN дуже велика, в таблиці 2 наведено лише кілька кращих варіантів структур, що продемонстрували високу точність класифікації.

Для навчання моделі використовувались набори даних DisProt та набір одноланцюгових послідовностей банку PDB, які сумарно склали 3807 послідовностей. В якості тестового набору використаний набір даних рентгенівської кристалографії банку PDB, що склав 10505 послідовностей. Кожен білок перетворювався в матрицю амінокислотних атрибутів, як показано на рисунку 2.4. В якості міток використані 4 класи, відповідно до стандарту класифікації внутрішньо невпорядкованих білків.

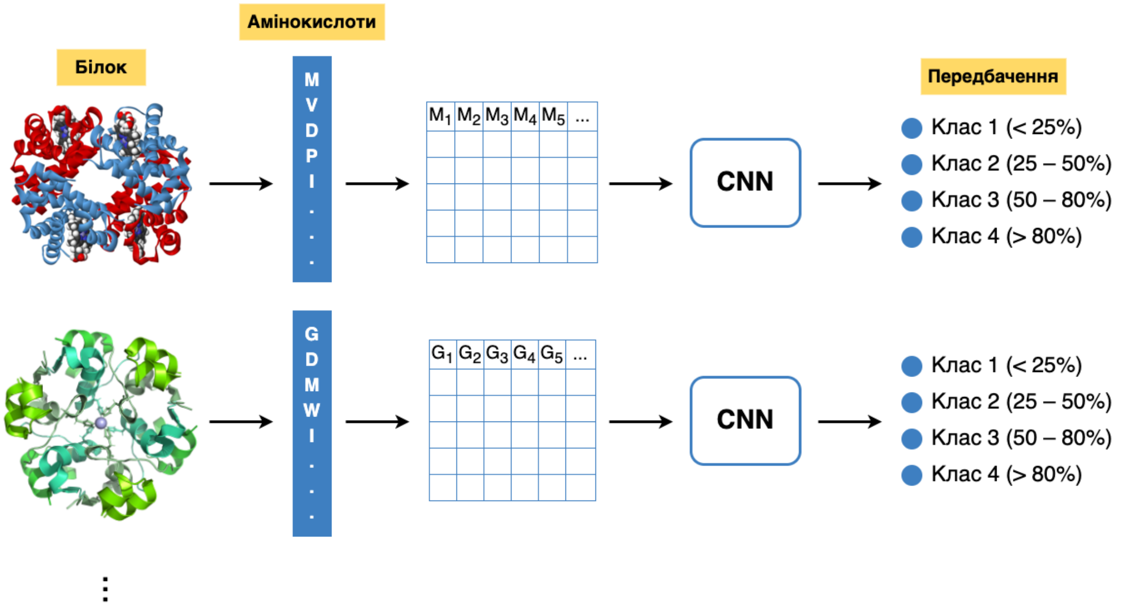

Рисунок 2.5 – Схема процесу роботи нейронної мережі.

Процес для кожного білка схематично зображений на рисунку 2.5. У CNN спочатку за допомогою кількох фільтрів отримувались карти ознак вхідних даних. Потім за допомогою шару субсемплінгу розмірність даних зменшувалась. Після одного або декількох згорткових шарів та субсемплінгу дані подавались на вхід повноз'єднаного шару, а значення класу отримувалось з виходу класифікатора. Точність кожної моделі оцінювалась як відношення загальної кількості правильно класифікованих послідовностей до загальної кількості тестових послідовностей.

Навчання проводилось з використанням GPU NVIDIA Tesla K80 у хмарному середовищі Google Colab.

2.2.4. Результати

Результати та параметри структур кращих з отриманих моделей наведені в таблиці 2.4.

Кількість згорткових шарів	Кількість фільтрів	Розмір повноз'єднаного шару	Розмір ядра фільтру	Час навчання, с	Точність
1	8	256	5	8966	89.65%
1	16	256	5	10698	90.82%
1	8	512	5	11878	92.03%
1	16	512	5	13021	92.87%
2	8,16	256	5	7763	91.62%
2	8,16	512	5	10446	93.24%
2	16,32	512	5	13855	90.19%

Таблиця 2.4 – Параметри та точність моделей CNN.

Як можна бачити, найкращу точність продемонструвала модель із двох згорткових шарів з 8 і 16 фільтрами, з розміром повноз'єднаного шару 512. Час навчання склав 10446 секунд.

При аналізі результатів було помічено, що при збільшенні кількості згорткових шарів час тренування не збільшувався, а навіть зменшувався.

Наприклад, час навчання моделі з 1 згортковим шаром з 8 фільтрами склав 8966 секунд, коли для навчання моделі з двома згортковими шарами з 8 та 16 фільтрами знадобилося лише 7763 секунди. Це відбувається завдяки субсемплінгу з максимізаційним агрегуванням при розмірі кластеру 2×2. За наявності лише одного згорткового шару, розмір матриці складав 60×60, але при двох згорткових шарах розмір матриці зменшувався до 30×30. Внаслідок зменшення розмірності даних час тренування моделі не зростав при додаванні більшої кількості згорткових шарів.

ВИСНОВКИ

Дана робота включала використання методів пошуку, підготовки, модифікації біологічних даних, статистичні методи обробки та аналізу, а також застосування спеціалізованого програмного забезпечення та інструментів машинного навчання.

Під час даного дослідження було відібрано та проаналізовано кращі з наявних у даний час алгоритми-предиктори для білків з внутрішньою невпорядкованістю. Кожен з них має індивідуальні особливості роботи та специфічні обмеження. Запропоновано метастратегію – поєднати результати індивідуальних предикторів з метою підвищення якості прогнозування. Методами статистичного аналізу було визначено 8 найбільш значущих індивідуальних предикторів, результати передбачення яких використані у якості вхідних даних для побудови метапредиктора з підвищеною ефективністю.

В результаті було побудовано модель метапредиктора на основі класифікатора Random Forest, яка має точність 92.21% для класифікації окремих амінокислотних залишків білка на приналежність до класу невпорядкованих.

На основі показників метапредиктора за допомогою AutoML бібліотеки AutoKeras було проведено пошук оптимальних гіперпараметрів CNN моделі-класифікатора повних білкових послідовностей. Проаналізовано моделі з різною кількістю згорткових шарів, фільтрів, з різними розмірами ядра фільтру та повноз'єднаного шару. Найкраща конфігурація моделі продемонструвала точність класифікації повної амінокислотної послідовності щодо рівня внутрішньої невпорядкованості білка на рівні 93%.

Подальші можливості покращення цього показника можуть включати більш тонку настройку інших параметрів нейронної мережі.

Список використаних джерел

1. Romero P., Obradovic Z., Kissinger C. B., Villafranca J. E., Guilliot S., Garner E., Dunker A. K. 1998. Thousands of proteins likely to have long disordered regions. Pac. Symp. Biocomput. 3: 437-448.
2. Dunker A. K., Obradovic Z., Romero P., Garner E. C., Brown C. J. 2000. Intrinsic protein disorder in complete genomes. Genome Inform. Ser. Workshop. Genome Inform. 11: 161-171.
3. Ward J. J., Sodhi J. S., McGuffin L. J., Buxton B. F., Jo- nes D. T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. J. Mol. Biol. 337: 635-645.
4. Oldfield C. J., Cheng Y., Cortese M. S., Romero P., Uvers- ky V. N., Dunker A. K. 2005. Coupled folding and binding with alpha-helix-forming molecular recognition elements. Biochemistry. 44: 12454-12 470.
5. Iakoucheva L. M., Brown C. J., Lawson J. D., Obradovic Z., Dunker A. K. 2002. Intrinsic disorder in cell-signaling and cancer-associated proteins. J. Mol. Biol. 323: 573-584.
6. Uversky V. N., Oldfield C. J., Dunker A. K. 2008. Intrinsically disordered proteins in human diseases: introducing the D2 concept. Annu. Rev. Biophys. 37: 215-246.
7. Cheng Y., LeGall T., Oldfield C. J., Dunker A. K., Uvers- ky V. N. 2006. Abundance of intrinsic disorder in protein associated with cardiovascular disease. Biochemistry. 45: 10448-10460.
8. Uversky V. N. 2008. Amyloidogenesis of natively unfolded proteins. Curr. Alzheimer Res. 5: 260-287.
9. Dunker A. K., Garner E., Guilliot S., Romero P., Albrecht K., Hart J., Obradovic Z., Kissinger C., Villafranca J. E. 1998. Protein disorder and the evolution of molecular recognition: theory, predictions and observations. Pac. Symp. Biocomput. 473-484.

10. Dunker A. K., Lawson J. D., Brown C. J., Williams R. M., Ro- mero P., Oh J. S., Oldfield C. J., Campen A. M., Ratliff C. M., Hipps K. W., Ausio J., Nissen M. S., Reeves R., Kang C., Kissin- ger C. R., Bailey R. W., Griswold M. D., Chiu W., Garner E. C., Obradovic Z. 2001. Intrinsically disordered protein. J. Mol. Graph. Model. 19: 26-59.

11. Dunker A. K., Obradovic Z. 2001. The protein trinity-linking function and disorder. Nature Biotechnol. 19: 805-806.

12. Dunker A. K., Cortese M. S., Romero P., Iakoucheva L. M., Uversky V. N. 2005. Flexible nets. The roles of intrinsic disorder in protein interaction networks. FEBS J. 272: 5129-5148.

13. Wright P. E., Dyson H. J. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J. Mol. Biol. 293: 321-331.

14. Uversky V. N., Gillespie J. R., Fink A. L. 2000. Why are «natively unfolded» proteins unstructured under physiologic conditions? Proteins. 41: 415-427.

15. Uversky V. N., Oldfield C. J., Dunker A. K. 2005. Showing your ID: intrinsic disorder as an ID for recognition, regulation and cell signaling. J. Mol. Recognit. 18: 343-384.

16. Tompa P. 2002. Intrinsically unstructured proteins. Trends Biochem. Sci. 27: 527-533.

17. Tompa P. 2005. The interplay between structure and function in intrinsically unstructured proteins. FEBS Lett. 579: 3346-3354.

18. Fink A. L. 2005. Natively unfolded proteins. Curr. Opin. Struct. Biol. 15: 35-41.

19. Borgia, A., Borgia, M.B., Bugge, K., Kissling, V.M., Heidarsson, P.O., Fernandes, C.B., Sottini, A., Soranno, A., Buholzer, K.J., Nettels, D., et al.: Extreme disorder in an ultrahigh-affinity protein complex. Nature 555(7694), 61–66 (2018)

20. Fischer, D.: 3d-shotgun: a novel, cooperative, fold-recognition meta-predictor. Proteins: Structure, Function, and Bioinformatics 51(3), 434–441 (2003)

21. Bi Zhao & Lukasz Kurgan (2021) Surveying over 100 predictors of intrinsic disorder in proteins, Expert Review of Proteomics, 18:12, 1019-1029, DOI: 10.1080/14789450.2021.2018304
22. Wan, J., Kang, S., Tang, C., Yan, J., Ren, Y., Liu, J., Gao, X., Banerjee, A., Ellis, L.B., Li, T.: Meta-prediction of phosphorylation sites with weighted voting and restricted grid search parameter selection. Nucleic acids research 36(4), 22–22 (2008)
23. Liu Y., Wang X., Liu B. A comprehensive review and comparison of existing computational methods for intrinsically disordered protein and region prediction. Brief Bioinform. 2019 Jan 18; 20(1): 330-346.
24. He B., Wang K., Liu Y., et al. Predicting intrinsic disorder in proteins: an overview. Cell Res. 2009 Aug; 19(8): 929-49.
25. Meng F., Uversky V., Kurgan L. Computational Prediction of Intrinsic Disorder in Proteins. Curr Protoc Protein Sci. 2017 Apr 3; 88: 2.16.1-2.16.14.
26. Deng X., Eickholt J., Cheng J. A comprehensive overview of computational protein disorder prediction methods. Mol Biosyst. 2012 Jan; 8(1): 114-21.
27. Li J., Feng Y., Wang X. et al. An Overview of Predictors for Intrinsically Disordered Proteins over 2010-2014. Int J Mol Sci. 2015 Sep 29; 16(10): 23446-62.
28. Pryor EE, Jr., Wiener MC. A critical evaluation of in silico methods for detection of membrane protein intrinsic disorder. Biophys J. 2014 Apr 15;106(8): 16, 38-49.
29. Dosztanyi Z., Meszaros B., Simon I. Bioinformatical approaches to characterize intrinsically disordered/unstructured proteins. Brief Bioinform. 2010 Mar; 11(2): 225-43.
30. Katuwawala A., Oldfield CJ, Kurgan L. Accuracy of protein-level disorder predictions. Brief Bioinform. 2020 Sep 25;21(5): 1509-1522.
31. Kurgan L., Li M., Li Y. The Methods and Tools for Intrinsic Disorder Prediction and their Application to Systems Medicine. In: Wolkenhauer O, editor. Systems Medicine. Oxford: Academic Press; 2021. p. 159-169.

32. Katuwawala A., Kurgan L. Comparative Assessment of Intrinsic Disorder Predictions with a Focus on Protein and Nucleic Acid-Binding Proteins. Biomolecules. 2020 Dec 4; 10(12).

33. Necci M., Piovesan D., Dosztanyi Z., et al. A comprehensive assessment of long intrinsic protein disorder from the DisProt database. Bioinformatics. 2018 Feb 1; 34(3): 445-452.

34. Peng Z.L., Kurgan L. Comprehensive comparative assessment of in-silico predictors of disordered regions. Curr Protein Pept Sci. 2012 Feb; 13(1): 6-18.

35. Necci M., Piovesan D., Predictors C., et al. Critical assessment of protein intrinsic disorder prediction. Nat Methods. 2021 May; 18(5): 472-481.

36. Jin Y., Dunbrack R.L., Jr. Assessment of disorder predictions in CASP6. Proteins. 2005; 61 Suppl 7: 167-75.

37. Bordoli L., Kiefer F., Schwede T. Assessment of disorder predictions in CASP7. Proteins. 2007; 69 Suppl. 8: 129-36.

38. Noivirt-Brik O., Prilusky J., Sussman J.L. Assessment of disorder predictions in CASP8. Proteins. 2009; 77 Suppl 9: 210-6.

39. Monastyrskyy B., Kryshtafovych A., Moult J., et al. Assessment of protein disorder region predictions in CASP10. Proteins. 2014 Feb; 82 Suppl 2: 127-37.

40. Melamud E., Moult J. Evaluation of disorder predictions in CASP5. Proteins. 2003; 53 Suppl 6: 561-5.

41. Monastyrskyy B., Fidelis K., Moult J., et al. Evaluation of disorder predictions in CASP9. Proteins. 2011; 79 Suppl 10: 107-18.

42. Lee, M.S., Oh, S.: Alternating decision tree algorithm for assessing protein interaction reliability. Vietnam Journal of Computer Science 1(3), 169–178 (2014)

43. Geng, H., Lu, T., Lin, X., Liu, Y., Yan, F.: Prediction of protein-protein interaction sites based on naive bayes classifier. Biochemistry research international (2015)

44. Qi, Y., Klein-Seetharaman, J., Bar-Joseph, Z.: Random forest similarity for protein-protein interaction prediction from multiple sources. In: Biocomputing 2005, pp. 531–542. World Scientific (2005)
45. Ramkumar, T., et al.: Analysis of multilayer perceptron machine learning approach in classifying protein secondary structures. Biomedical Research (2016)
46. Williams, D., Hinton, G.: Learning representations by back-propagating errors. Nature 323(6088), 533–538 (1986)
47. Li, M., Cho, S.B., Ryu, K.H.: A novel approach for predicting disordered regions in a protein sequence. Osong public health and research perspectives 5(4), 211–218 (2014)
48. Zeng, H., Edwards, M.D., Liu, G., Gifford, D.K.: Convolutional neural network architectures for predicting dna–protein binding. Bioinformatics 32(12), 121–127 (2016)
49. LeCun Y. et al. 2015. Deep learning. Nature, 521, 436-444.
50. Alipanahi B. et al. 2015. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. Nat. Biotechnol., 33, 831-838.
51. Zhou J., Troyanskaya O.G. 2015. Predicting effects of noncoding variants with deep learning-based sequence model. Nat. Methods, 12, 931-934.
52. Prilusky, J., Felder, C.E., Zeev-Ben-Mordehai, T., Rydberg, E.H., Man, O., Beckmann, J.S., Silman, I., Sussman, J.L.: Foldindex©: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21(16), 3435–3438 (2005)
53. Linding, R., Jensen, L.J., Diella, F., Bork, P., Gibson, T.J., Russell, R.B.: Protein disorder prediction: implications for structural proteomics. Structure 11(11), 1453–1459 (2003)
54. Ulijn, R.V., Lampel, A.: Order/disorder in protein and peptide-based biomaterials. Israel Journal of Chemistry (2019)
55. Sussman, J.L., Lin, D., Jiang, J., Manning, N.O., Prilusky, J., Ritter, O., Abola, E.: Protein data bank (pdb): database of three-dimensional structural information

of biological macromolecules. Acta Crystallographica Section D: Biological Crystallography 54(6), 1078–1084 (1998)

56. Sickmeier, M., Hamilton, J.A., LeGall, T., Vacic, V., Cortese, M.S., Tantos, A., Szabo, B., Tompa, P., Chen, J., Uversky, V.N., et al.: Disprot: the database of disordered proteins. Nucleic acids research 35(suppl 1), 786–793 (2006)
57. Guex, N., Peitsch, M.C.: Swiss-model and the swiss-pdb viewer: an environment for comparative protein modeling. electrophoresis 18(15), 2714–2723 (1997)
58. Yang, Z.R., Thomson, R., McNeil, P., Esnouf, R.M.: Ronn: the bio-basis function neural network technique applied to the detection of natively disordered regions in proteins. Bioinformatics 21(16), 3369–3376 (2005)
59. Walsh, I., Martin, A.J., Di Domenico, T., Tosatto, S.C.: Espritz: accurate and fast prediction of protein disorder. Bioinformatics 28(4), 503–509 (2011)
60. Hu G, Katuwawala A, Wang K, et al. flDPnn: Accurate intrinsic disorder prediction with putative propensities of disorder functions. Nat Commun 2021; 12(1): 4438.
61. Hanson J., Paliwal K.K., Litfin T., et al. SPOT-Disorder 2: improved protein intrinsic disorder prediction by ensembled deep learning. Genom. Proteom. Bioinform. 2019; 17(6): 645–56.
62. Mirabello C., Wallner B. rawMSA: end-to-end deep learning using raw multiple sequence alignments. PLoS ONE 2019; 14(8): e0220182.
63. Wang S., Ma J., Xu J. AUCpreD: proteome-level protein disorder prediction by AUC-maximized deep convolutional neural fields. Bioinformatics 2016;32 (17):i672–9.
64. Chengbin Hu, Yiru Qin, Chuan Ye et al. A High Accurate Machine Learning Meta-Strategy for the Prediction of Intrinsically Disorder Proteins, 2021, PREPRINT (Version 1), DOI: 10.21203/rs.3.rs-903129/v1
65. Bhatt D, Patel C, Talsania H, Patel J, Vaghela R, Pandya S, Modi K, Ghayvat H. CNN Variants for Computer Vision: History, Architecture, Application, Challenges and Future Scope. Electronics. 2021; 10(20):2470. https://doi.org/10.3390/electronics10202470

66. Van Der Lee, R., Buljan, M., Lang, B., Weatheritt, R.J., Daughdrill, G.W., Dunker, A.K., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D.T., et al.: Classification of intrinsically disordered regions and proteins. Chemical reviews 114(13), 6589–6631 (2014)