國立臺灣海洋大學資訊工程學系
碩士學位論文考試

預測超級細菌綠膿桿菌蛋白質間的交互作用
Predicting Protein-Protein Interactions of Superbug Pseudomonas Aeruginosa

指導教授：阮議聰 博士
報 告 人 ： 張 祐 琪
報告日期：2024/6/19

目錄·CONTENTS

01 研究背景

02 研究動機

03 系統架構

- ✓ 測試資料集
- ✓ 偽造非交互作用蛋白質方法
- ✓ 蛋白質描述方法
- ✓ 協方差(AC)
- ✓ 簡化胺基酸(RAAA)

04 機器學習

- ✓ KNN
- ✓ SVM
- ✓ 深度學習
- ✓ DNN
- ✓ Drop out層

05 實驗與數據分析

- ✓ 評估方法
- ✓ 系統比較

06 結論

PART-01

研　究　背　景

研　究　背　景

蛋　白　質

是由一條或多條胺基酸殘基的長鏈組成的一個大型的生物分子(Large Biomolecule)或高分子(Macromolecule)功能：酶催化、信號傳導、免疫防禦等。
龐大且複雜的分子，由數百或數千個20種胺基酸(Amino Acid)組合形成胺基酸序列

交 互 作 用

指兩種或以上的蛋白質結合的過程，通常是為了執行其生化功能。在細胞中，大量蛋白質元件組成分子機器，透過蛋白質交互作用執行細胞內多數重要的分子過程，如 DNA 複製。在生物體中，蛋白質會彼此結合在一起，形成複合體，執行特殊的功能。

預 測 系 統

序列取得容易 : 蛋白質一級結構胺基酸序列可由蛋白質基因的核苷酸(Nucleotide)序列推得。
降低實驗成本 : 公開網站工具或資料庫等等提供使用，不必準備實驗儀器及花費大量時間等待結果。

PART-02

研　究　動　機

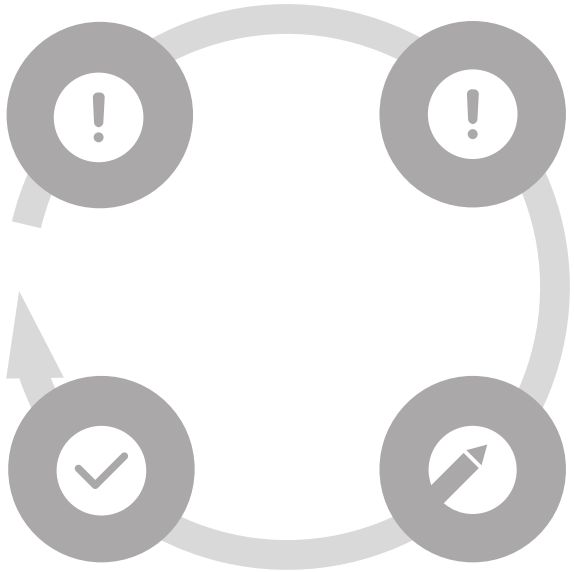
研　究　動　機

超級細菌

指的是細菌有多重抗生素的抗藥性。這樣的超級細菌可能對於市面上多數的抗生素都產生有抗藥性，甚至對所有的抗生素都有抗藥性，以致於感染的病患會面臨沒有藥物可以治療的狀況。

動機目標

透過不同蛋白質描述方法和不同偽造非交互作用蛋白質對方法，再使用SVM、DNN、KNN的機器學習預測系統，找到當中最好的預測系統組合。



早期研究高成本耗時

早期研究方法：免疫沉澱法、染色質免疫沉澱方法等。
在實驗預測上面有他們各自優點，但同時存在大量樣品、過程繁瑣、耗費時間的缺點。

系統生物學

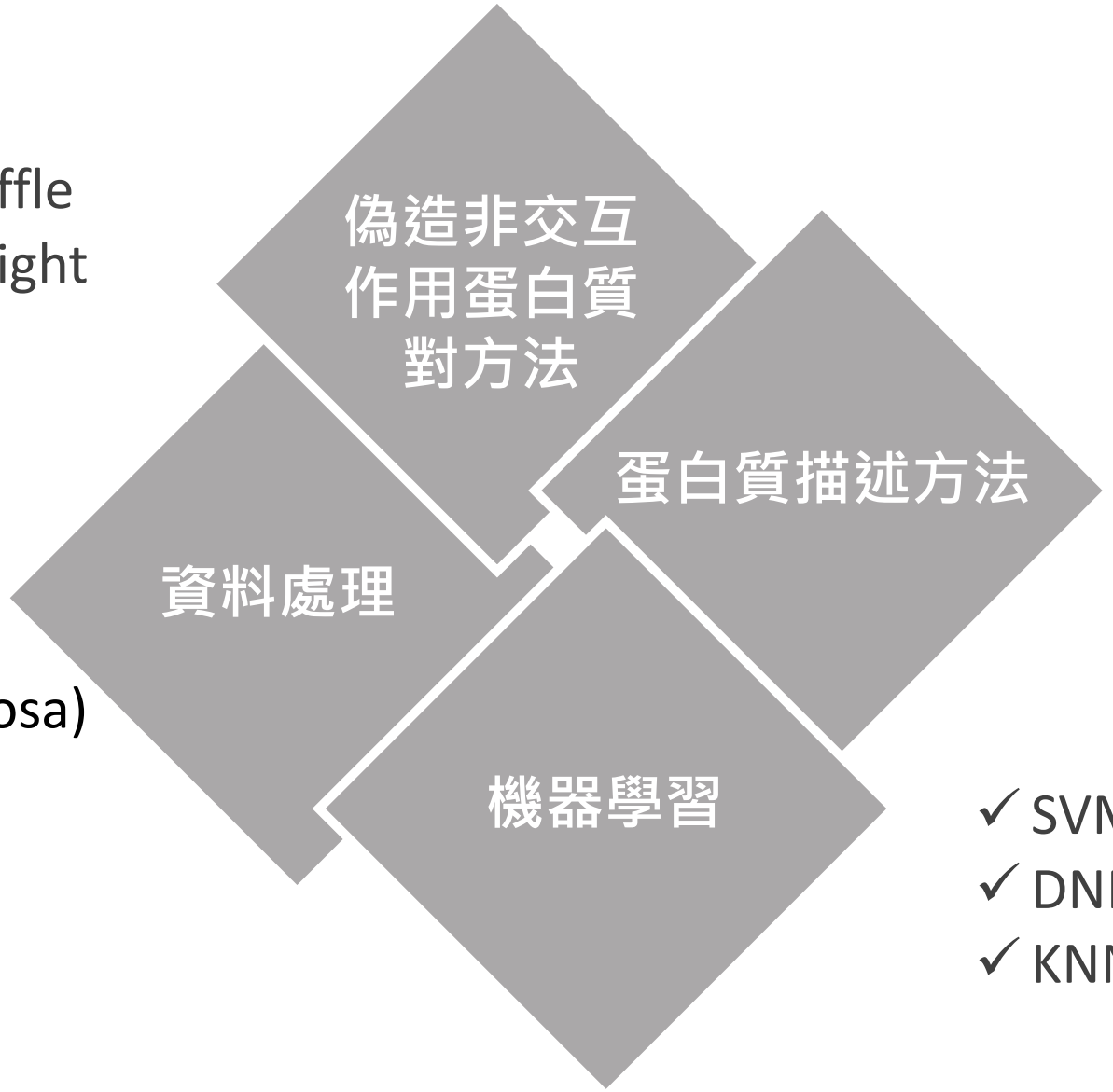
不同於以往只注意個別的基因和蛋白質的實驗生物學，研究所有的基因、所有的蛋白質和組分間的所有相互關係；其目標是：對複雜的生物系統構建出計算的數學模型，從總體上預測生物系統的真實性。

PART-03

系　統　架　構

系　統　架　構

超級細菌 - 綠膿桿菌
(Pseudomonas Aeruginosa)



- ✓ Ushuffle
- ✓ LeftRight

- ✓ 簡化胺基酸(RAAA)
- ✓ 協方差(AC)

- ✓ SVM
- ✓ DNN
- ✓ KNN

系　統　架　構

蛋白質交互作用資料集

STRING資料庫內的綠膿桿菌(Pseudomonas Aeruginosa)，也被列為一種超級細菌。

訓練集和測試集處理

- 刪除蛋白質胺基酸序列長度小於50
- 刪除重複的蛋白質交互作用對
- 刪除訓練集和測試集共同的蛋白質對
- 刪除相似蛋白質(cdhit40)

偽造非交互作用蛋白質對方法

> 本實驗方法名稱 ： Ushuffle
>
> 將一個蛋白質的胺基酸序列維持原來序列長度下亂序排列視為新的蛋白質，同時與他配對的也是偽造不交互作用對。

> 本實驗方法名稱 ： LeftRight
>
> 將蛋白質交互作用對的集合中重新隨機配對成偽造不交互作用的蛋白質對並保證唯一。

蛋白質描述方法 - 標準胺基酸組成(Amino Acid Composition, AAC)

將蛋白質序列中20種標準胺基酸的出現次數除以蛋白質序列總長，即可求得20個特徵向量的組成百分比之數值陣列。準確率不高卻極具代表性。

$$f_i = \frac{\text{Total number of amino acoid } i}{L}, where\ f_i (i = 1, 2, \dots, 20)$$

$$P = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ f_i \\ \vdots \\ f_{20} \end{bmatrix}$$

蛋白質描述方法 - N-肽組成(N-peptide Composition, N-peptide)

將蛋白質序列中每一種N肽的出現次數NP_i ，除以所有N肽在蛋白質序列上所有可能組合的總個數，求得N肽在蛋白質序列上組成的百分比。陣列長度(特徵向量維度)為20^{N} (N為肽長度)。

舉例 : N=1，會形成A、C、D……Y，共20個特徵向量

N=2，會形成AA、AC、AD……YY，共400個特徵向量

$$f_i = \frac{NP_i}{T} \qquad N_P = \begin{bmatrix} f_1 \\ f_2 \\ \vdots \\ \vdots \\ f_{20^n} \end{bmatrix}, where\ f_{20^n}(1 \le n \le L)$$

蛋白質描述方法 - 簡化胺基酸 (Reduced Amino Acid Alphabets, RAAA)

20種標準胺基酸中有些胺基酸的物理化學特性是相似或相同的，因此20種標準胺基酸可以被分類成更少種的胺基酸，不僅簡化了系統的複雜度，也保留了蛋白質區域結構訊息的能力，本論文所使用的為 PT20、CP13、CP11，前三種的簡化胺基酸，在後面實驗分別以 RAAA1、RAAA2、RAAA3 代稱。

舉例：

PT20，長度2的肽(peptide)

$\Rightarrow 20^2 = 400$維

⇒ 一條簡化胺基酸 = 400維

⇒ 兩條簡化胺基酸 = 800維

Cluster profiles	簡化胺基酸RAAA / N=2 (在長度2肽底下的維度)
PT20	400
CP13	169
CP11	121

蛋白質描述方法 -協方差(Auto covariance, AC)

在描述蛋白質序列中，我們會用協方差(Auto covariance, AC)來描述說一條蛋白質序列殘基和他的鄰居之間的關係。協方差主要被用來將蛋白質序列的數值特徵（如物理化學性質）轉換成特徵向量，便於機器學習模型處理。

- ✓ 數值表示: 將胺基酸轉換為數值特徵，如疏水性、體積、極性等。
- ✓ 計算協方差: 計算相鄰胺基酸特徵之間的協方差，反映胺基酸對間距為lag的關係。
- ✓ 特徵向量: 將不同間距的協方差值組合成統一的特徵向量。

PART-04

機 器 學 習

K-近鄰演算法(K-Nearest Neighbors Algorithm,K-NN)

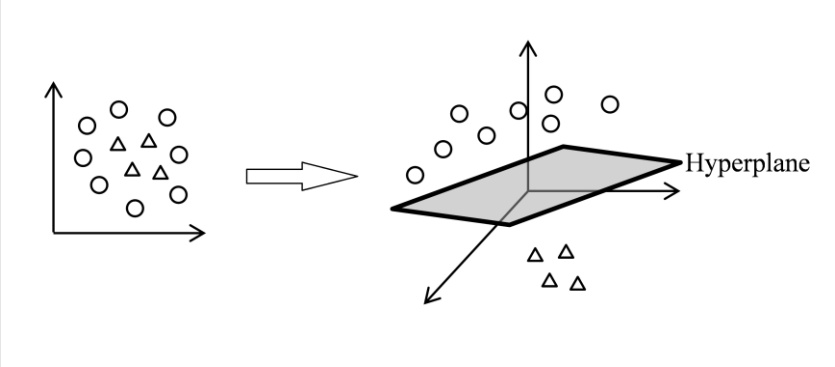
一種演算法在監督式(Supervised Learning)下可被用來解決分類和回歸問題，採用測量不同特徵值之間的距離方法進行分類，一個測試資料的分類是由其K個鄰居所屬的族群多數決而定，適合多分類問題，也因為多數決投票，K值為奇數比較不會有平局問題。

優點：簡單易理解、實現簡單、無需訓練。
缺點：計算效率比較低、類別不平衡問題。

支持向量機(Support Vector Machine, SVM)

支持向量機概念：

將數據映射到高維的空間裡，建立一個最大分隔數據的超平面(Hyperplane)，使兩邊分類盡可能與超平面的距離最大化。目標為，找出一個或多個超平面，並且作為分類依據。



支持向量機泛化效能好可以做分類與迴歸分析，也可解決線性與非線性問題，使SVM被廣泛應用，例如：影像辨識、文字分類和生物科技等。

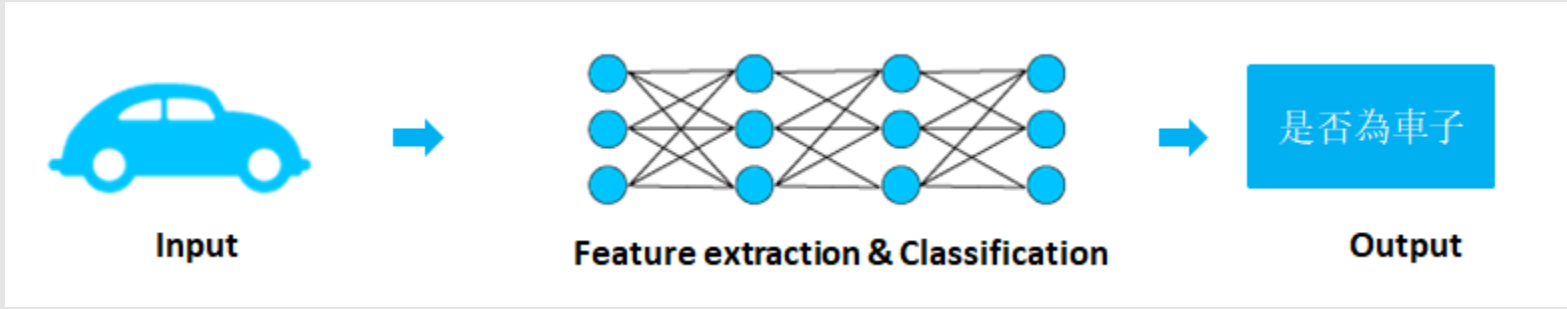
深度學習(Deep Learning , DL)

深度學習原理，訓練資料透過建立的訓練模型不斷優化，最後完成訓練模型，再將我們的測試資料輸入比對，測出準確性。

人工神經網路架構

例：深度神經網路(DNN)、卷積神經網路(CNN)以及遞迴神經網路(RNN)

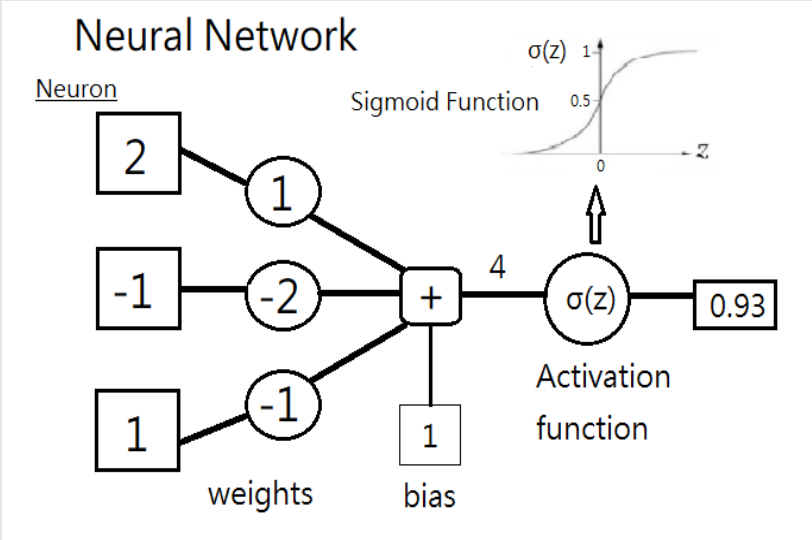
一個經典的神經網路包含三個層次，1.輸入層，2.隱藏層，3.輸出層

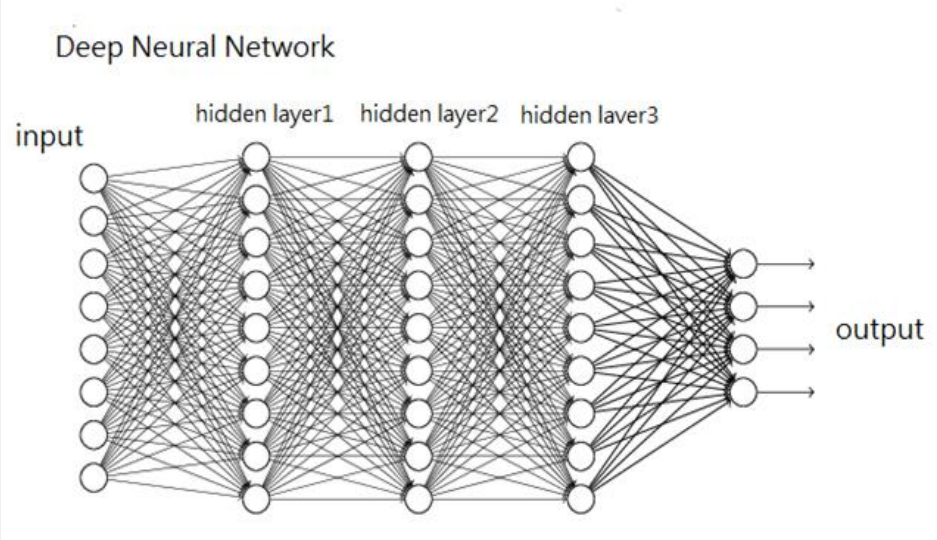


深度神經網路(Deep Neural Network , DNN)

神經網路訓練是透過多個神經元(Neurons)組成，而一個神經元的基本組成是輸入的變數、權重(Weight)、偏差(Bias)以及激勵函數(Activation function)。是一個多個神經元且多層隱藏層(Hidden Layer)的神經網路訓練。

理想的正確分類得分值與目前的權重所計算出的得分值之間的差距被稱為損失函數(loss)，減小loss值，重複迭代。

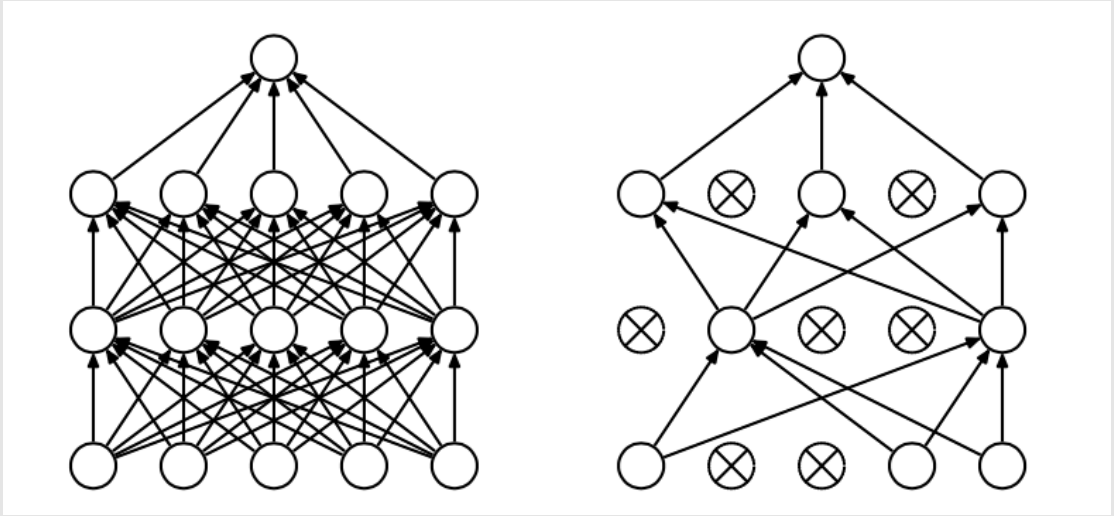




Drop Out 層

在機器學習中，若是模型的引數太多，且訓練樣本過少，就容易產生過擬合的現象，許多機器學習的模型都會遇到該問題，為了解決該問題就會在模型當中加入 Drop Out 以有效的緩解過擬和的發生。

Drop Out 可以做為訓練的一種 trick 供使用，在每個訓練中都會忽略掉一半的特徵檢測，讓每個特徵檢測器不會過於互相依賴，以至於整個模型不會過於依賴某些區域的特徵。



PART-05

實驗與數據分析

評估方式

在機器學習的分類預測問題中，一般較為常用的評斷方法是使用準確率(Accuracy)，對於給定的測試資料集，分類器正確分類的樣本數與總樣本數比。因為使用準確率比較容易評斷整體預測系統的預測表現，但在本篇論文中，我們主要會使用精確率(Precision)以及準確率(Accuracy)來評估分類器的預測表現，靈敏度(Sensitivity)作為輔助 。

準確率(Accuracy) = $\frac{TP+TN}{TP+TN+FP+FN}$

精確率(Precision) = $\frac{TP}{TP+FP}$

靈敏度(Sensitivity) = $\frac{TP}{TP+FN}$

	實際Positive	**實際Negative**
預測Positive	TP (True Positive)	FP (False Positive)
預測Negative	FN (False Negative)	TN (True Negative)

實驗結果

交互作用蛋白質對(Positive data): 訓練集有11289筆、測試集有2995筆。
非交互作用蛋白質對(Negative data): 訓練集的Ushuffle-K1有11289筆、LeftRight的22578筆，測試集的Ushuffle-K1有2995筆、LeftRight有5990筆。

蛋白質描述方法:

- ✓ 協方差(AC)，並設定lag值為10、20、30、40、50。
- ✓ 簡化胺基酸(RAAA)，並搭配不同長度肽，RAAA1_lt2、RAAA2_lt2、RAAA2_lt3、RAAA3_lt3。

再分別使用三種分類器KNN、SVM以及深度神經網路(DNN)，DNN會再分為有加入Drop out層和沒加入的(DNN_ND)來做我們預測交互作用的實驗。

實驗結果

Ushuffle-seed0287_peptide_nm1_RAAA1_lt2

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.850	0.851	0.850
SVM	2617	399	2596	378	0.874	0.868	0.870
DNN	2719	105	2890	276	0.908	**0.963**	0.936
DNN_ND	2862	153	2842	133	0.956	0.949	**0.952**

實驗結果

Ushuffle-seed0287_peptide_nm1_RAAA2_lt2

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.812	0.818	0.812
SVM	2624	370	2625	371	0.876	0.876	0.876
DNN	2721	193	2802	274	0.909	**0.934**	**0.922**
DNN_ND	2682	221	2774	313	0.895	0.924	0.911

實驗結果

Ushuffle-seed0287_peptide_nm1_RAAA2_lt3

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.827	0.834	0.827
SVM	2719	215	2780	276	0.908	0.927	0.918
DNN	2854	127	2868	141	0.953	**0.957**	0.955
DNN_ND	2864	134	2861	131	0.956	0.955	**0.956**

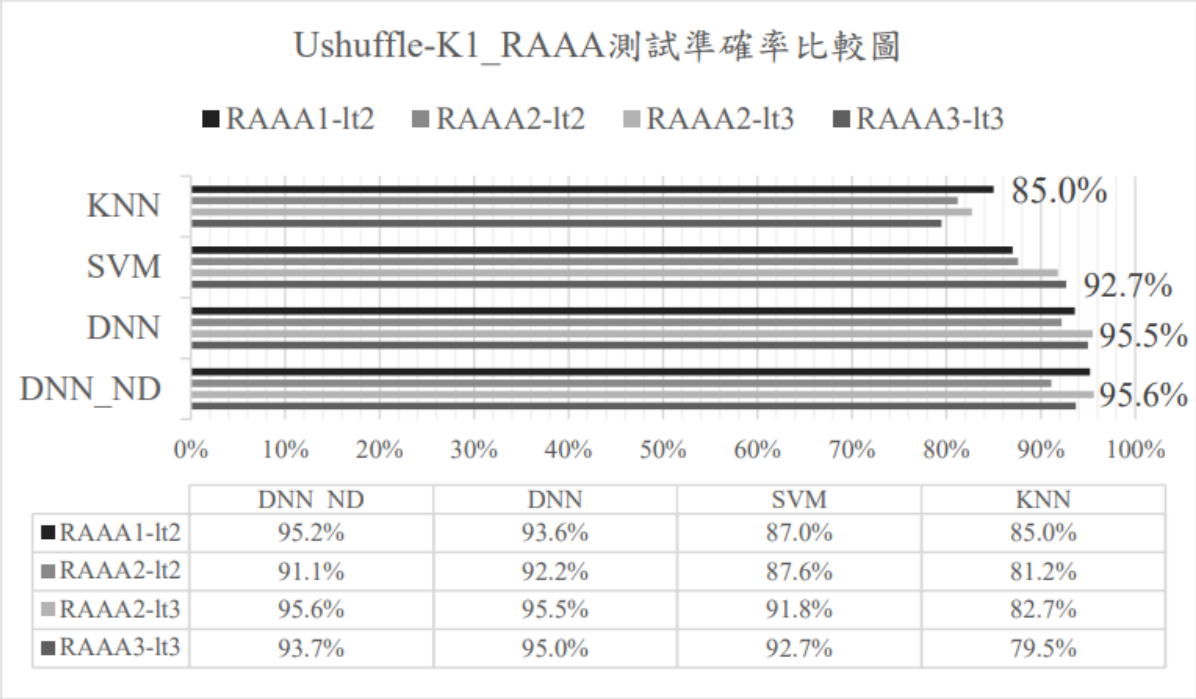
實驗結果

Ushuffle-seed0287_peptide_nm1_RAAA3_lt3

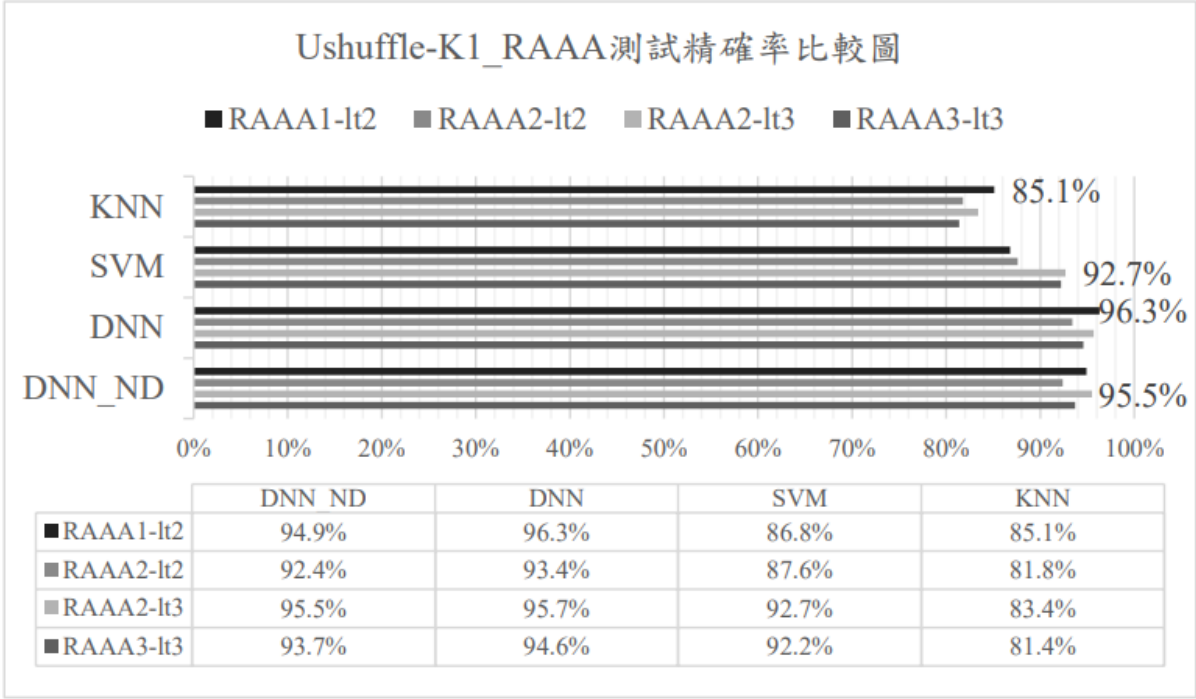
	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.795	0.814	0.795
SVM	2791	236	2759	204	0.932	0.922	0.927
DNN	2858	164	2831	137	0.954	**0.946**	**0.950**
DNN_ND	2806	190	2805	189	0.937	0.937	0.937

實驗與數據分析

實驗結果



	DNN_ND	DNN	SVM	KNN
RAAA1-lt2	95.2%	93.6%	87.0%	85.0%
RAAA2-lt2	91.1%	92.2%	87.6%	81.2%
RAAA2-lt3	95.6%	95.5%	91.8%	82.7%
RAAA3-lt3	93.7%	95.0%	92.7%	79.5%



	DNN_ND	DNN	SVM	KNN
RAAA1-lt2	94.9%	96.3%	86.8%	85.1%
RAAA2-lt2	92.4%	93.4%	87.6%	81.8%
RAAA2-lt3	95.5%	95.7%	92.7%	83.4%
RAAA3-lt3	93.7%	94.6%	92.2%	81.4%

實驗結果

LeftRight-seed0-287_peptide_nm1_RAAA1_lt2

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.603	**0.609**	0.659
SVM	1346	946	5044	1649	0.449	0.587	0.711
DNN	1838	1282	4708	1157	0.614	0.589	**0.729**
DNN_ND	1611	1219	4771	1384	0.538	0.569	0.710

實驗結果

LeftRight-seed0-287_peptide_nm1_RAAA2_lt2

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.608	**0.614**	0.663
SVM	912	713	5277	2083	0.305	0.561	0.689
DNN	1753	1399	4591	1242	0.585	0.556	**0.706**
DNN_ND	1224	1070	4920	1771	0.409	0.534	0.684

實驗結果

LeftRight-seed0-287_peptide_nm1_RAAA2_lt3

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.590	0.599	0.654
SVM	963	912	5078	2032	0.322	0.514	0.672
DNN	1655	1227	4763	1340	0.553	0.574	**0.714**
DNN_ND	1191	765	5225	1804	0.398	**0.609**	**0.714**

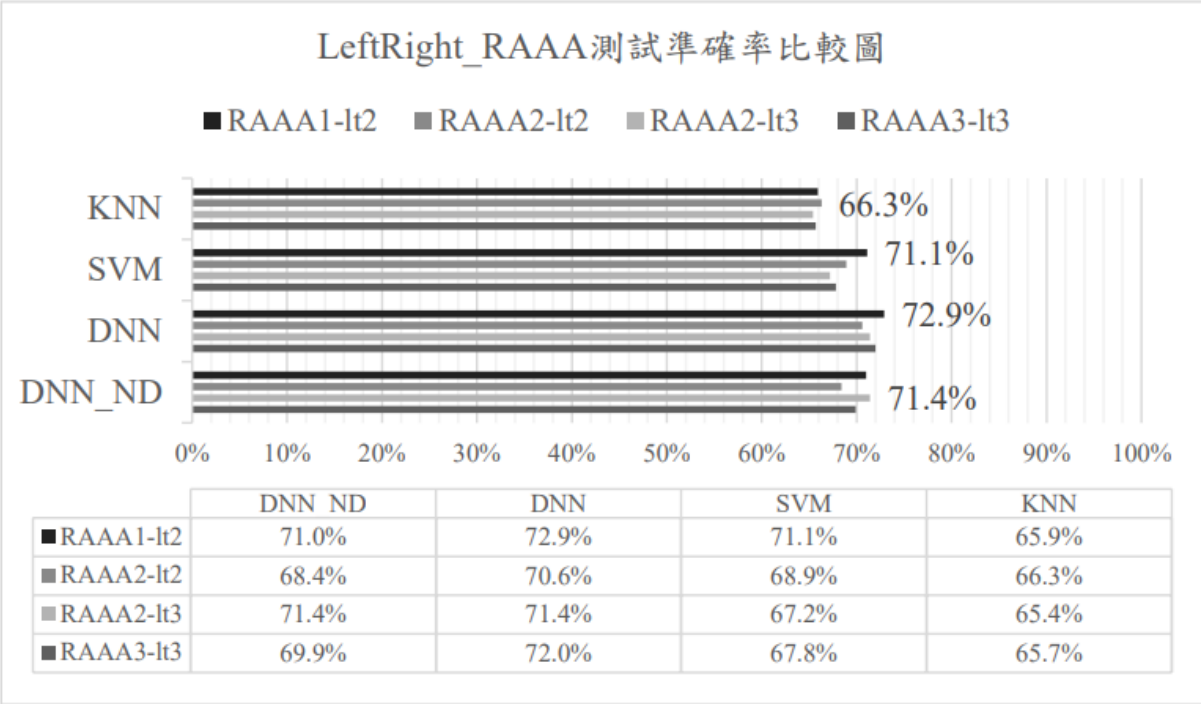
實驗結果

LeftRight-seed0-287_peptide_nm1_RAAA3_lt3

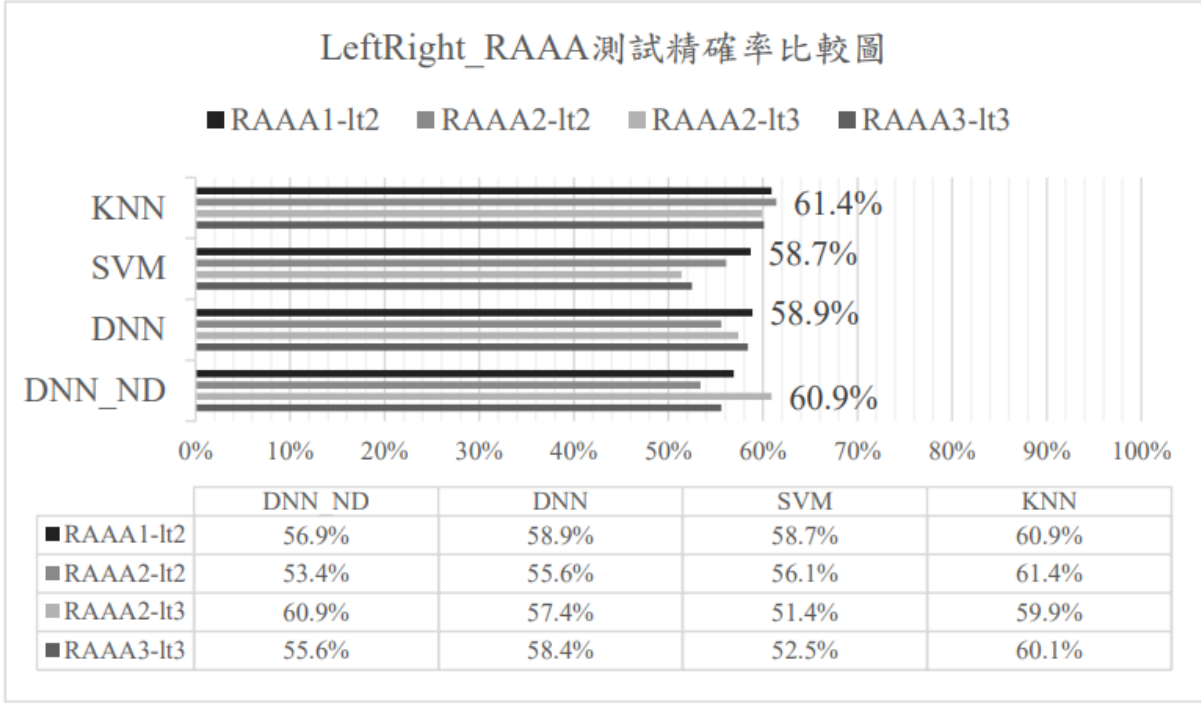
	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.587	**0.601**	0.657
SVM	1078	976	5014	1917	0.360	0.525	0.678
DNN	1659	1181	4809	1336	0.554	0.584	**0.720**
DNN_ND	1476	1181	4809	1519	0.493	0.556	0.699

實驗與數據分析

實驗結果



	DNN_ND	DNN	SVM	KNN
RAAA1-lt2	71.0%	72.9%	71.1%	65.9%
RAAA2-lt2	68.4%	70.6%	68.9%	66.3%
RAAA2-lt3	71.4%	71.4%	67.2%	65.4%
RAAA3-lt3	69.9%	72.0%	67.8%	65.7%



	DNN_ND	DNN	SVM	KNN
RAAA1-lt2	56.9%	58.9%	58.7%	60.9%
RAAA2-lt2	53.4%	55.6%	56.1%	61.4%
RAAA2-lt3	60.9%	57.4%	51.4%	59.9%
RAAA3-lt3	55.6%	58.4%	52.5%	60.1%

實驗結果

Ushuffle-seed0-K1-287_AC_10

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.815	0.820	0.815
SVM	2693	312	2683	302	0.899	0.896	0.898
DNN	2730	307	2688	265	0.912	**0.899**	**0.905**
DNN_ND	2715	330	2665	280	0.907	0.892	0.898

實驗結果

Ushuffle-seed0-K1-287_AC_20

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.799	0.812	0.799
SVM	2609	407	2588	386	0.871	0.865	0.868
DNN	2711	343	2652	284	0.905	**0.888**	**0.895**
DNN_ND	2657	386	2609	338	0.887	0.873	0.879

實驗結果

Ushuffle-seed0-K1-287_AC_30

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.804	0.809	0.804
SVM	2783	233	2762	212	0.929	**0.923**	**0.926**
DNN	2741	313	2682	254	0.915	0.898	0.905
DNN_ND	2628	374	2621	367	0.877	0.875	0.876

實驗結果

Ushuffle-seed0-K1-287_AC_40

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.807	0.815	0.807
SVM	2798	254	2741	197	0.934	**0.917**	**0.925**
DNN	2735	286	2709	260	0.913	0.905	0.909
DNN_ND	2654	332	2663	341	0.886	0.889	0.888

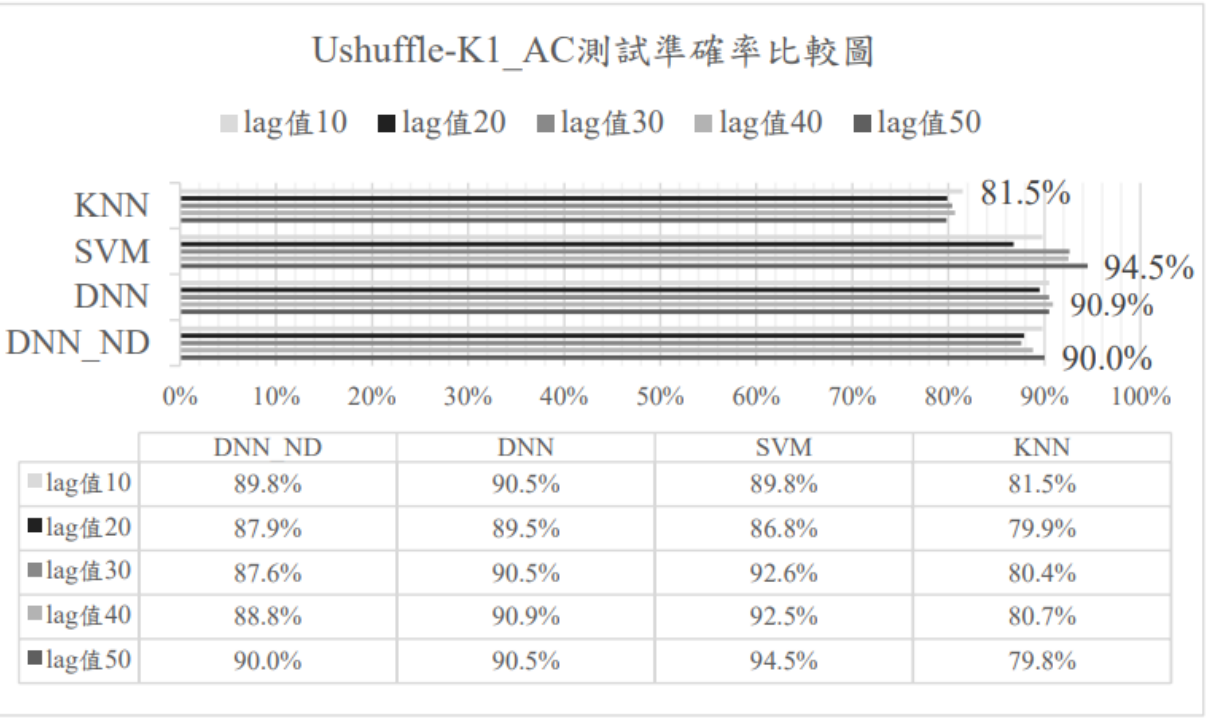
實驗結果

Ushuffle-seed0-K1-287_AC_50

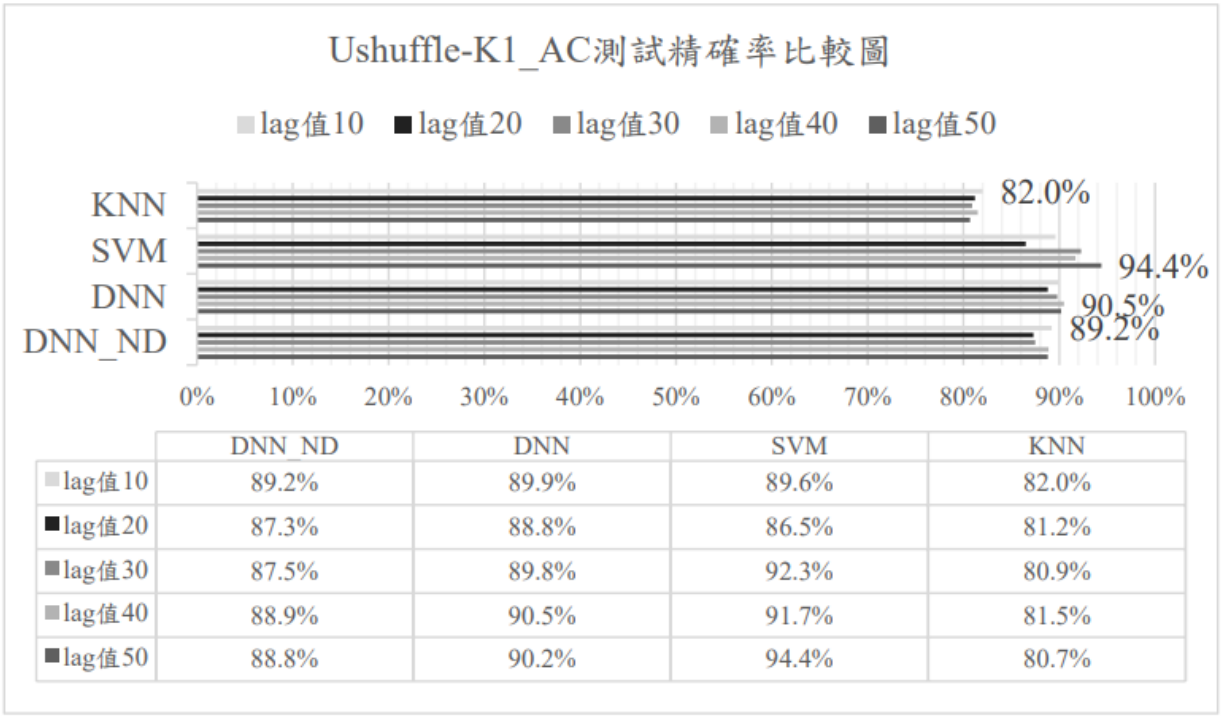
	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.798	0.807	0.798
SVM	2832	169	2826	163	0.946	**0.944**	**0.945**
DNN	2719	294	2701	276	0.908	0.902	0.905
DNN_ND	2739	345	2650	256	0.915	0.888	0.900

實驗與數據分析

實驗結果



	DNN_ND	DNN	SVM	KNN
lag值10	89.8%	90.5%	89.8%	81.5%
lag值20	87.9%	89.5%	86.8%	79.9%
lag值30	87.6%	90.5%	92.6%	80.4%
lag值40	88.8%	90.9%	92.5%	80.7%
lag值50	90.0%	90.5%	94.5%	79.8%



	DNN_ND	DNN	SVM	KNN
lag值10	89.2%	89.9%	89.6%	82.0%
lag值20	87.3%	88.8%	86.5%	81.2%
lag值30	87.5%	89.8%	92.3%	80.9%
lag值40	88.9%	90.5%	91.7%	81.5%
lag值50	88.8%	90.2%	94.4%	80.7%

實驗結果

LeftRight-287_AC_10

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.575	**0.600**	0.658
SVM	772	715	5275	2223	0.258	0.519	**0.673**
DNN	1035	1299	4691	1960	0.346	0.443	0.637
DNN_ND	950	1083	4907	2045	0.317	0.467	0.652

實驗結果

LeftRight-287_AC_20

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.574	**0.591**	0.653
SVM	945	799	5191	2050	0.316	0.542	**0.683**
DNN	916	1082	4908	2079	0.306	0.458	0.648
DNN_ND	1098	1230	4760	1897	0.367	0.472	0.652

實驗結果

LeftRight-287_AC_30

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.583	**0.601**	0.660
SVM	887	777	5213	2108	0.296	0.533	**0.679**
DNN	1138	1308	4682	1857	0.380	0.465	0.648
DNN_ND	1196	1342	4648	1799	0.399	0.471	0.650

實驗結果

LeftRight-287_AC_40

	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.577	**0.600**	0.661
SVM	1140	1047	4943	1855	0.381	0.521	**0.677**
DNN	1161	1281	4709	1834	0.388	0.475	0.653
DNN_ND	1050	1107	4883	1945	0.351	0.487	0.660

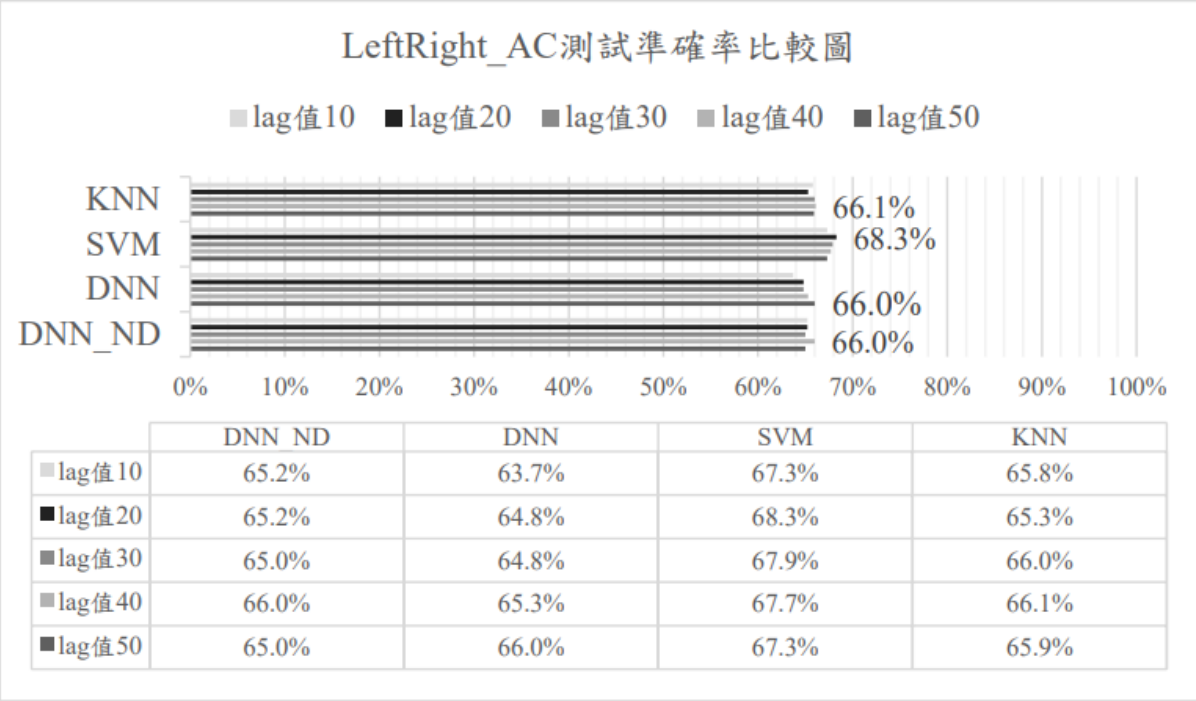
實驗結果

LeftRight-287_AC_50

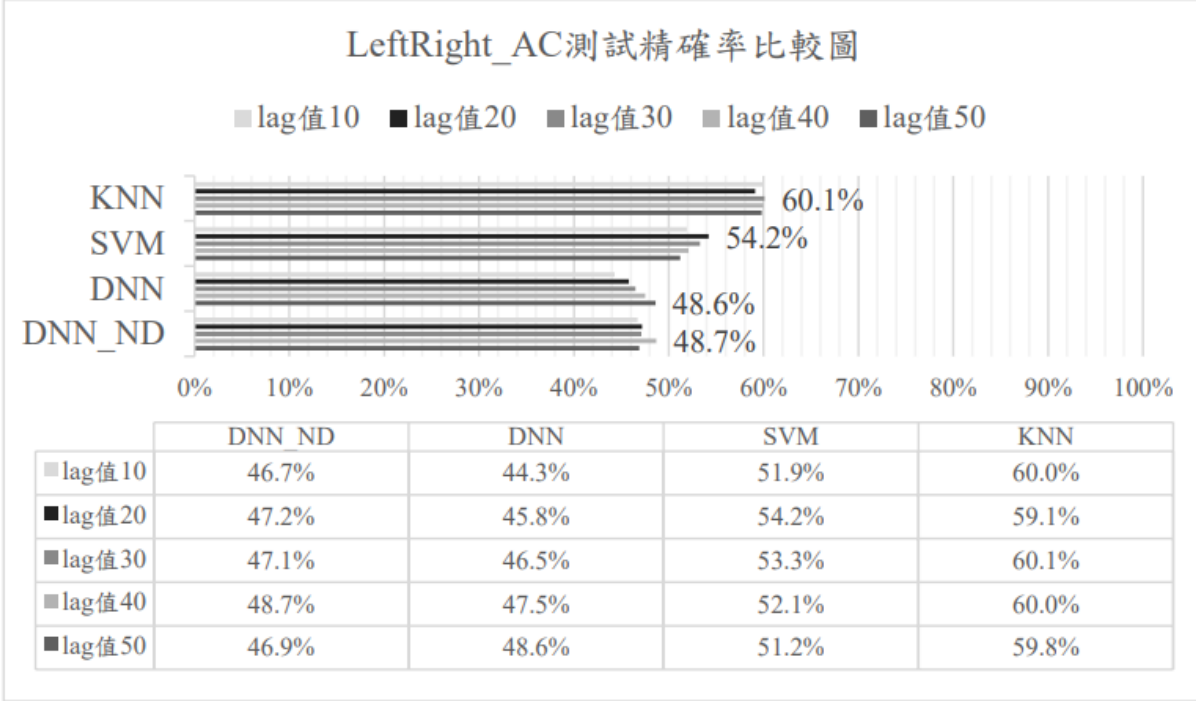
	TP	FP	TN	FN	Sensitivity	Precision	Test ACC
KNN					0.579	**0.598**	0.659
SVM	1162	1108	4882	1834	0.388	0.512	**0.673**
DNN	1038	1097	4893	1957	0.347	0.486	0.660
DNN_ND	1161	1315	4675	1834	0.388	0.469	0.650

實驗與數據分析

實驗結果



	DNN_ND	DNN	SVM	KNN
lag值10	65.2%	63.7%	67.3%	65.8%
lag值20	65.2%	64.8%	68.3%	65.3%
lag值30	65.0%	64.8%	67.9%	66.0%
lag值40	66.0%	65.3%	67.7%	66.1%
lag值50	65.0%	66.0%	67.3%	65.9%



	DNN_ND	DNN	SVM	KNN
lag值10	46.7%	44.3%	51.9%	60.0%
lag值20	47.2%	45.8%	54.2%	59.1%
lag值30	47.1%	46.5%	53.3%	60.1%
lag值40	48.7%	47.5%	52.1%	60.0%
lag值50	46.9%	48.6%	51.2%	59.8%

實驗結果

各分類器在Ushuffle方法下測試準確率最高

Classfier	偽造蛋白質不交互作用發法	蛋白質描述方法	測試準確率
KNN	Ushuffle	RAAA1-lt2	85.0%
SVM	Ushuffle	AC(lag=10)	94.5%
DNN	Ushuffle	RAAA2-lt3	95.5%
DNN_ND	Ushuffle	RAAA2-lt3	**95.6%**

實驗結果

各分類器在LeftRight方法下測試準確率最高

Classfier	偽造蛋白質不交互作用方法	蛋白質描述方法	測試準確率
KNN	LeftRight	RAAA2-lt2	66.3%
SVM	LeftRight	RAAA1-lt2	71.1%
DNN	LeftRight	RAAA1-lt2	**72.9%**
DNN_ND	LeftRight	RAAA2-lt3	71.4%

實驗結果

各分類器在Ushuffle方法下測試精確率最高

Classfier	偽造蛋白質不交互作用發法	蛋白質描述方法	測試準確率
KNN	Ushuffle	RAAA1-lt2	85.1%
SVM	Ushuffle	AC(lag=50)	94.4%
DNN	Ushuffle	RAAA1-lt2	**96.3%**
DNN_ND	Ushuffle	RAAA2-lt3	95.5%

實驗結果

各分類器在LeftRight方法下測試精確率最高

Classfier	偽造蛋白質不交互作用方法	蛋白質描述方法	測試準確率
KNN	LeftRight	RAAA2-lt2	**61.4%**
SVM	LeftRight	RAAA1-lt2	58.7%
DNN	LeftRight	RAAA1-lt2	58.9%
DNN_ND	LeftRight	RAAA2-lt3	60.9%

PART-06

結　　論

結論

實驗結果顯示出兩個**蛋白質描述方法**，簡化胺基酸(RAAA)相比協方差(AC)有高一點的準確率和精確率，但差異不會太大。在**偽造蛋白質不交互作用對的方法**上，Ushuffle相比LeftRight有較高的準確率和精確率。分類器上，在DNN模型中有加入DropOut層(DNN)，雖然一般來說可以減緩過擬和的發生，但就實驗數據看下來也沒有更好的感覺，論文所使用的DNN模型還有可以改進的地方。

結論

在使用**RAAA作為蛋白質描述方法時**，DNN的系統測試準確性效果相對比較好，但使用**協方差(AC)作為蛋白質描述方法時**，SVM的系統測試準確性效果會比較好，而KNN準確性不管在哪種蛋白質描述方法幾乎都略輸於SVM和DNN。從精確率的角度來看，**在Ushuffle方法下**，一樣是DNN和SVM優於KNN，KNN大部分效果都很差，但**在LeftRight方法下**，KNN的效果會是最好的。整體來說DNN和SVM在本次實驗結果的表現是比較好的，KNN還有再調整進步的空間。未來可以加入更多種類的偽造蛋白質不交互作用方法以及蛋白質描述方法來進行比較。

最終實驗結果，**最高準確率**是在使用 Ushuffle 偽造非交互作用蛋白質對方法，蛋白質描述方法 RAAA2-lt3 及機器學習模型 DNN_ND 的 **95.6%**。**最高精確率**是在使用 Ushuffle 偽造非交互作用蛋白質對方法，蛋白質描述方法 RAAA1-lt2 及機器學習模型 DNN 的 **96.3%**。

感謝大家觀看

報告人：張祐琪