Afiahayati, S.Kom., M.Cs., Ph.D.

E-mail : afia@ugm.ac.id

Research interests : Machine Learning, Bioinformatics, Artificial Intelligence

Scopus ID/ Google Scholar ID : https://www.scopus.com/authid/detail.uri?authorId=55940697500 / https://scholar.google.com/citations?user=GU_FTN0AAAAJ&hl=en&oi=ao

EDUCATION

2011 – 2015	Ph.D. from Dept. of Biosciences and Informatics, Keio University, Japan
2008 - 2010	M.Cs. from Dept. of Computer Science and Electronics, Universitas Gadjah Mada
2004 - 2008	S.Kom. from Dept. of Computer Science and Electronics, Universitas Gadjah Mada

TEACHING EXPERIENCES (selected)

Undergraduate and Graduate Program	Machine Learning, Deep Learning, Bioinformatics, Artificial Intelligence, Computational Intelligence

EXPERIENCES (selected)

2022 – present	Head of Computer Laboratory, Fac. of Mathematics and Natural Sciences, Universitas Gadjah Mada
2011 – present	Lecturer, Dept. of Computer Science and Electronics, Fac. of Mathematics and Natural Sciences, Universitas Gadjah Mada
2019 – 2021	Head of Intelligent System Laboratory, Fac. of Mathematics and Natural Sciences, Universitas Gadjah Mada
2020 (3 months)	Visiting Researcher, Keio University, Japan
2015 (1 month)	Short course of Genome Informatics, Joint Genome Institute, University of California, Berkeley, US
2013 – 2015	Schlumberger Faculty for The Future (FFTF) Fellowship
2010	Short course of Computational Logic, held by TU Dresden, Germany

PUBLICATIONS (selected)

2021 **Afiahayati**, Yap Bee Wah, Yunita Sari, Diyah Utami Kusumaning Putri, Aina Musdholifah, Retantyo Wardoyo, Forecasting the cumulative COVID-19 cases in Indonesia using machine learning, *PeerJ Computer Science*, in review

2019 **Afiahayati**, Anarossi, E., Yanuaryska, R.D., Nuha, F.U., Mulyana, S., Comet assay classification for buccal Mucosa's DNA damage measurement with super tiny dataset using transfer learning, *Intelligent Information and Database Systems: Recent Developments* 830, 279-289

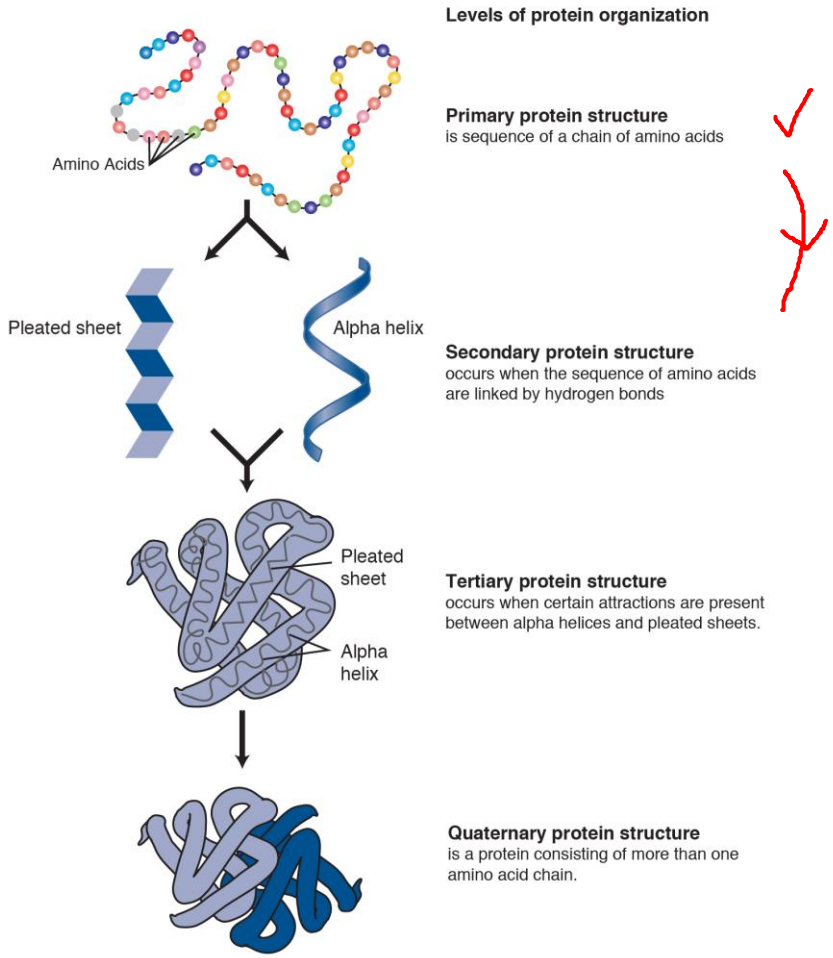
2015 **Afiahayati**, Sato, K., Sakakibara, Y., MetaVelvet-SL: an extension of the Velvet assembler to a de novo metagenomic assembler utilizing supervised learning, *DNA Research 22(1)*, 69-77

UNIVERSITAS
GADJAH MADA

Prediksi Struktur Sekunder Protein Menggunakan *Deep Learning*

Afiahayati
Departemen Ilmu Komputer dan Elektronika
Universitas Gadjah Mada
afia@ugm.ac.id

Struktur Protein



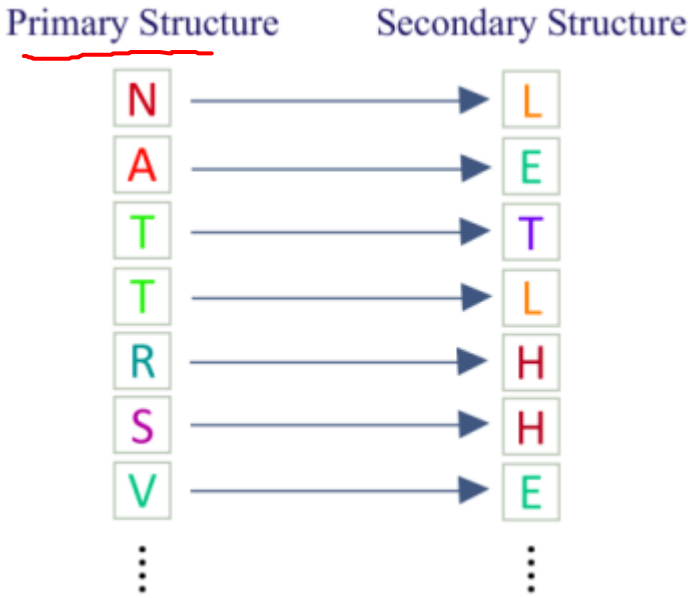
Courtesy: National Human Genome Research Institute

Fungsi Protein dapat dilihat dari strukturnya

Semakin tinggi strukturnya maka semakin mudah untuk menentukan fungsi biologis dari protein tersebut

Struktur Protein

Symbol	Abbreviation	Name
A	Ala	Alanina
C	Cys	Cysteine
D	Asp	Aspartic Acid
E	Glu	Glutamic Acid
F	Phe	Phenylalanine
G	Gly	Glycine
H	His	Histidine
I	Ile	Isoleucine
K	Lys	Lysine
L	Leu	Leucine
M	Met	Methionine
N	Asn	Asparagine
P	Pro	Proline
Q	Gln	Glutamine
R	Arg	Arginine
S	Ser	Serine
T	Thr	Threonine
V	Val	Valine
W	Trp	Tryptophan
Y	Tyr	Tyrosine



8 Class Name	8 Class Symbol	3 Class Name	3 Class Symbol
α-helix 3_{10}-helix π-helix	H G I	*Helix*	H
B-Strand B-bridge	E B	*Sheet*	E
Loop/Irreguler B-Turn Bend	L T S	*Coil / Loop*	C

Mengapa Memprediksi Struktur Protein?

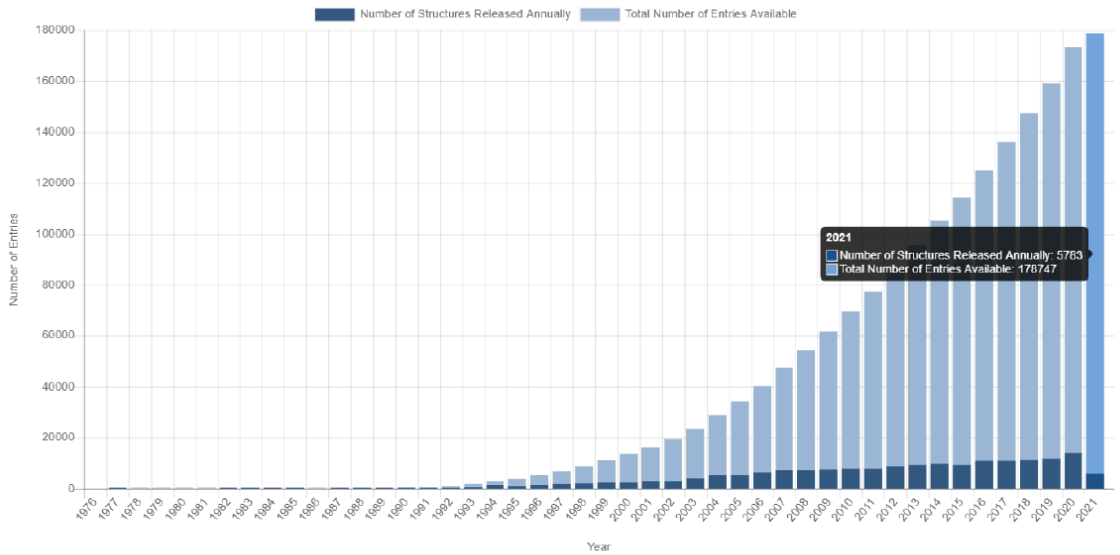
UNIVERSITAS GADJAH MADA

X-Ray Crystallography dan Nuclear Magnetic Resonance Spectroscopy (NMR) -> **waktu yang sangat banyak** dan **biaya yang mahal**

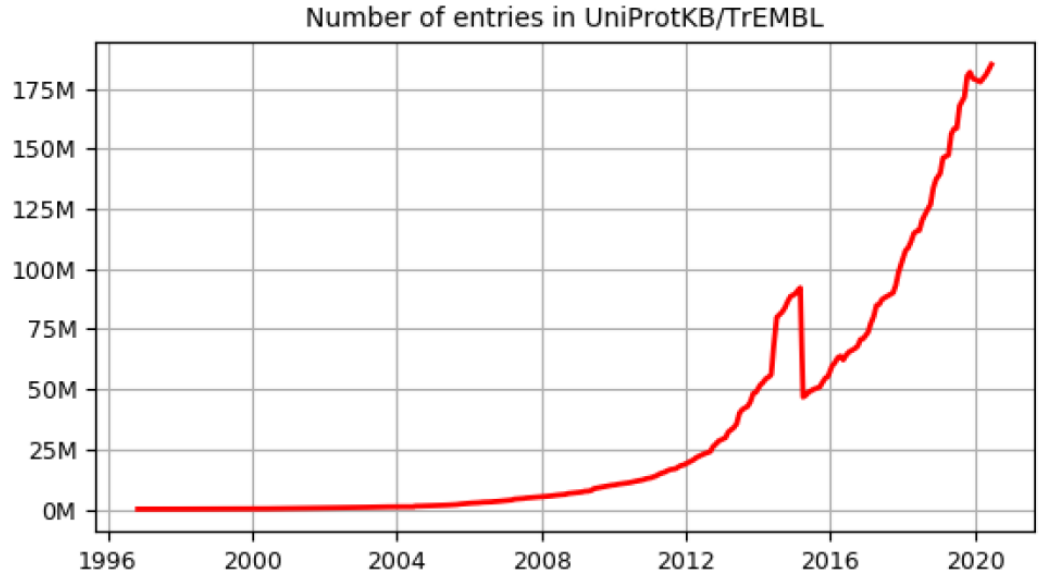
(Metfessel dan Saurugger (1993), Long dan Tian (2019)).

Solusi : Menggunakan pendekatan komputasi.

Mengapa Memprediksi Struktur Protein?



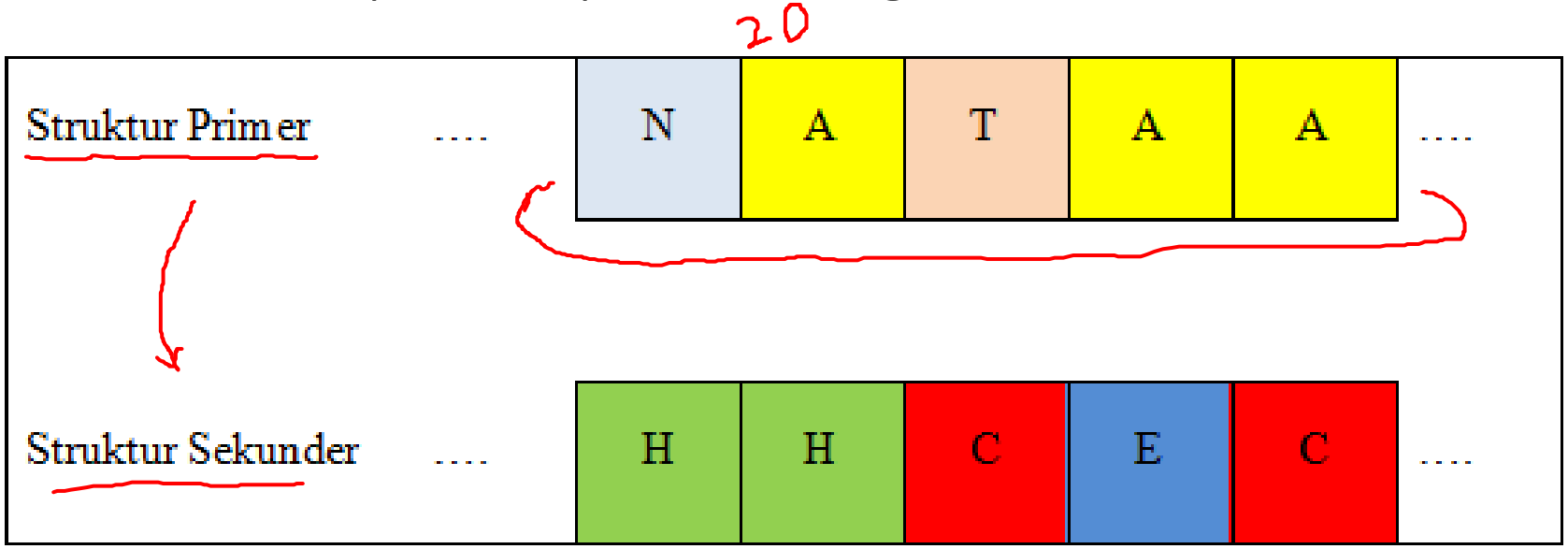
Gambar 1.2 Jumlah Struktur Protein di PDB tahun 2021 (PDB, 2021)



Gambar 1.3 Jumlah Sekuens Protein di UniProt Database tahun 2021 (UniProt, 2021)

Prediksi Struktur Sekunder Protein

Pendekatan Komputasi: Sequence Labelling



3 Kelas dan 8 Kelas

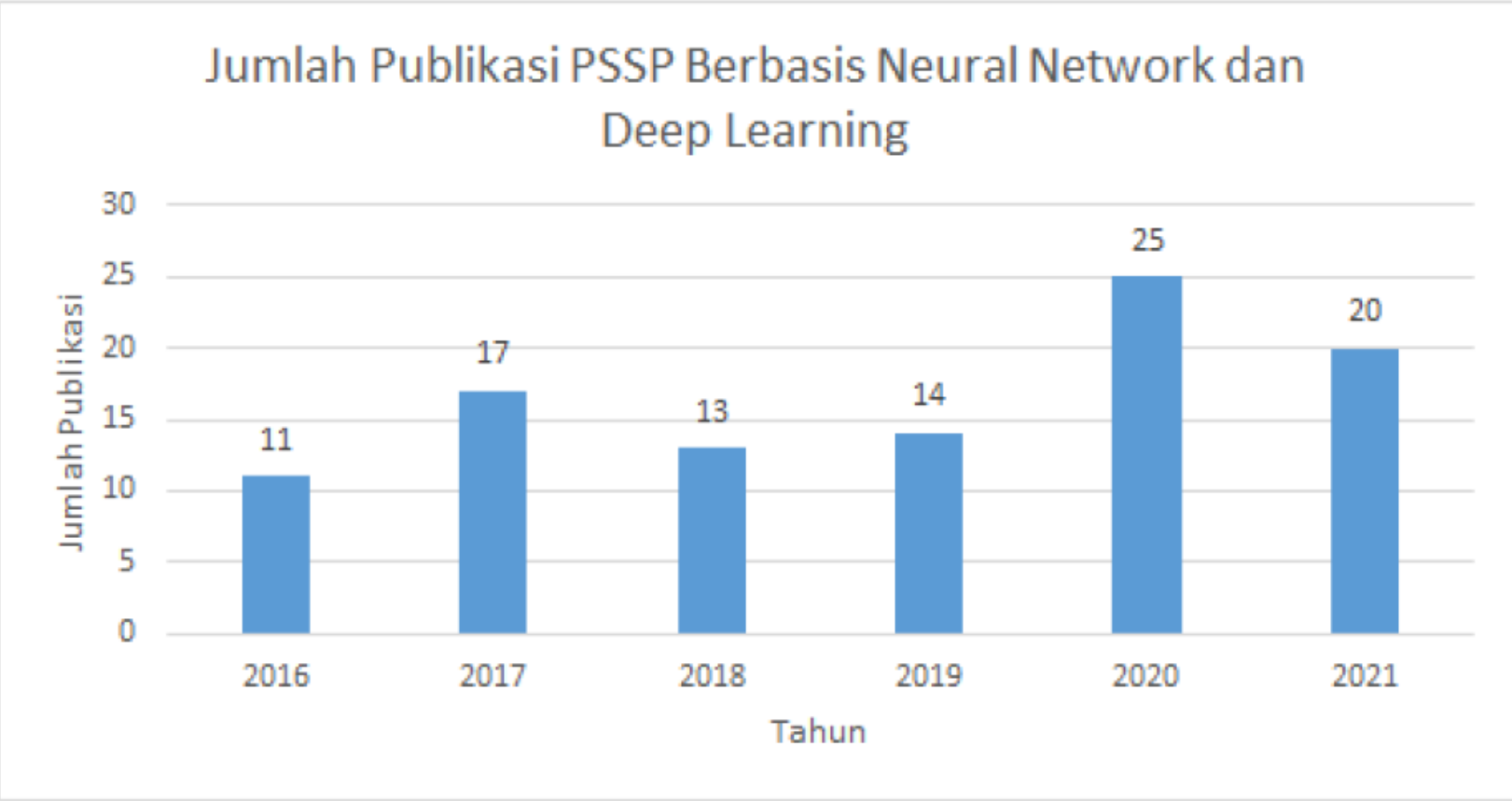
Kelas DSSP	Simbol 8 Kelas	Simbol 3 Kelas	Nama Kelas
3_{10}-helix α-helix π-helix	G H I	H	*Helix*
B-strand	E	E	*Sheet*
Isolated B-bridge Bend Turn Lainnya	B S T -	C	*Coil / Loop*

Tingkat akurasi diukur dengan persamaan **Q3** dan **Q8**

$$Q3 = \frac{\sum_{i=K} P_{ii}}{\sum_{i=K} \sum_{j=K} P_{ij}}, K \in \{H, E, C\}$$

$$Q8 = \frac{\sum_{i=K} P_{ii}}{\sum_{i=K} \sum_{j=K} P_{ij}}, K \in \{L, B, E, I, S, T, H, G\}$$

Publikasi PSSP berbasis NN dan DL



Publikasi PSSP dengan Neural Network dan Deep Learning pada Tahun 2016-2021 (Ismi, 2021)

PSSP Model Architecture

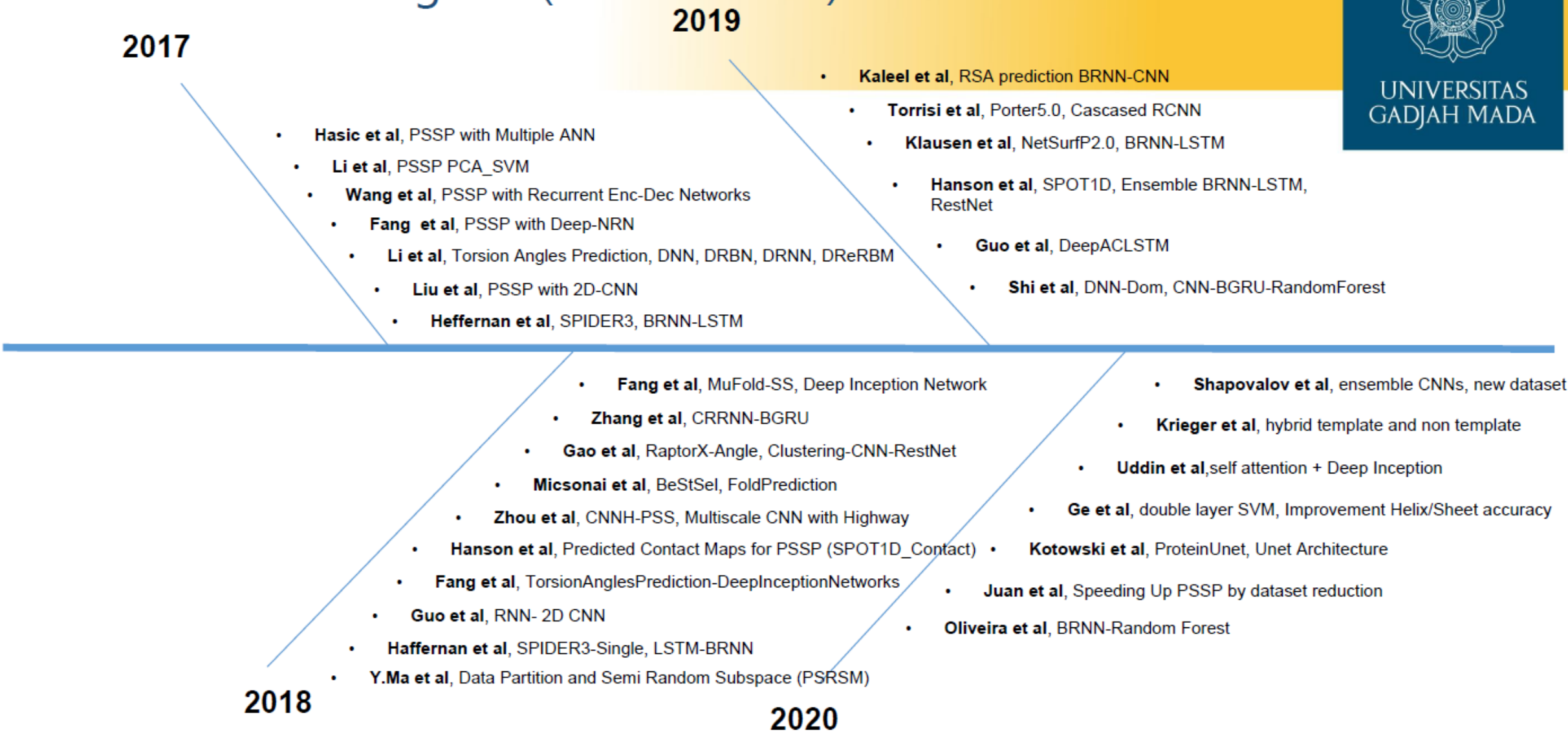
Akurasi tertinggi pada PSSP saat ini diperoleh dari arsitektur Deep Learning.

Beberapa model deep learning yang digunakan pada PSSP saat ini:

1. CNN
2. RNN-LSTM
3. GRU
4. Inception Network/ Capsule Network

(Ismi, 2021)

Fish Bone Diagram (PSSP & PSFP)



(Ismi, 2021)

UNIVERSITAS
GADJAH MADA

Predicting Secondary Structure of Protein Using Hybrid of Convolutional Neural Network and Support Vector Machine

International Journal of Intelligent Engineering & Systems
Vol.14, No.1, 2021
(Scopus Q3)

Vincent Michael Sutanto[1], Zaki Indra Sukma[2], Afiahayati[1*]

[1]Department of Computer Science and Electronics, Universitas Gadjah Mada, Indonesia
[2]Business Intelligence Data Engineering Division, Gojek, Indonesia
Corresponding author : afia@ugm.ac.id
(http://www.inass.org/2021/2021022823.pdf)

Dataset

CullPDB (Zhou dan Troyanskaya, 2014)

- 5365 sekuens asam amino yang tersimpan dalam bentuk array berukuran 5365 x 700 x 57
- digunakan sebagai data latih

CB513

- 514 sekuens asam amino yang tersimpan dalam bentuk array berukuran 514 x 700 x 57
- digunakan sebagai data uji

Dataset

Indeks	Panjang	Keterangan	Fungsi	Digunakan
0-21	22	*Orthogonal Input Profile*	Fitur	Ya
22-30	9	Label struktur sekunder (8 kelas)	Label	Ya
31	1	N - Terminal	-	Tidak
32	1	C - Terminal	-	Tidak
33-34	2	*Solvent Accessibility*	-	Tidak
35-56	22	*PSSM Profile*	Fitur	Ya

Position Specific Scoring Matrix

UNIVERSITAS GADJAH MADA

- PSSM digunakan untuk mengetahui peluang atau skor kemunculan suatu asam amino pada suatu posisi dalam sekuens asam amino.
- PSSM juga menunjukkan seberapa besar kemungkinan suatu asam amino pada suatu posisi berubah menjadi asam amino lain tanpa mengubah fungsi dari protein.
- Skor PSSM yang tinggi mengindikasikan suatu asam amino akan berubah menjadi asam amino lainnya.

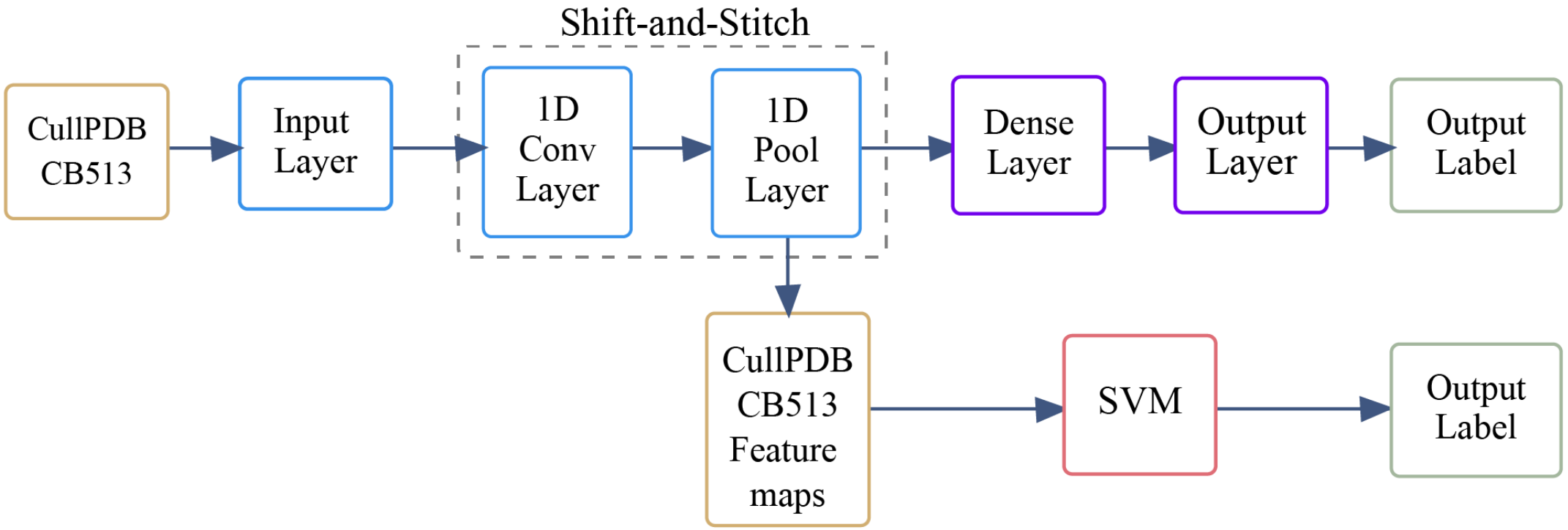
Metode: Kombinasi CNN + SVM

UNIVERSITAS GADJAH MADA

1-Dimensional CNN akan digunakan dalam menangkap fitur spasial dan memperkaya fitur dari sekuens asam amino

SVM akan digunakan memprediksi struktur sekunder protein berdasarkan fitur dari CNN

Metode: Kombinasi CNN + SVM

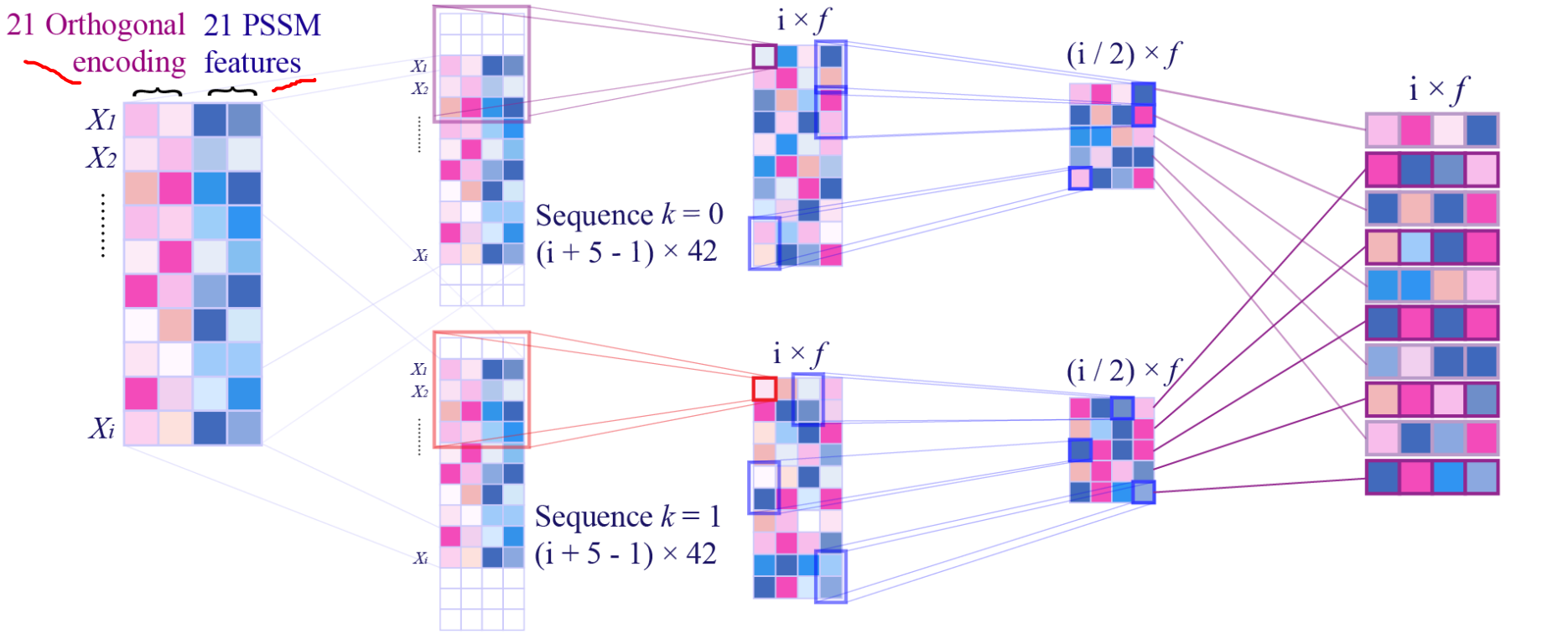


Shift-and-Stitch CNN

Shift and stitch adalah cara yang digunakan oleh Lin, et al. (2016) untuk mengatasi permasalahan berkurangnya panjang sekuens yang disebabkan oleh operasi konvolusi dan operasi pooling.

Dibagi kedalam 2 tahapan: **shift** dan **stitch**

Shift-and-Stitch CNN



"Shift" on original sequence (i × 42)

convolution layers with convolution size $w = 5$

Max pooling layers

"Stitch" on shifted sequence

Shift-and-stitch sequence (i × f)

Shift-and-Stitch CNN

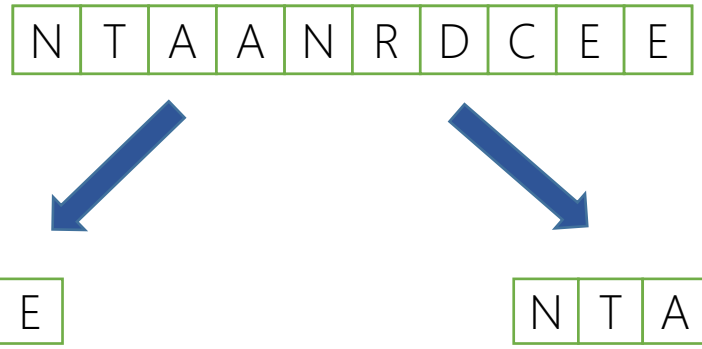
CONV SIZE = 5
MAXPOOL SIZE = 2

N	T	A	A	N	R	D	C	E	E

Panjang Awal = 10

Shift-and-Stitch CNN

CONV SIZE = 5
MAXPOOL SIZE = 2

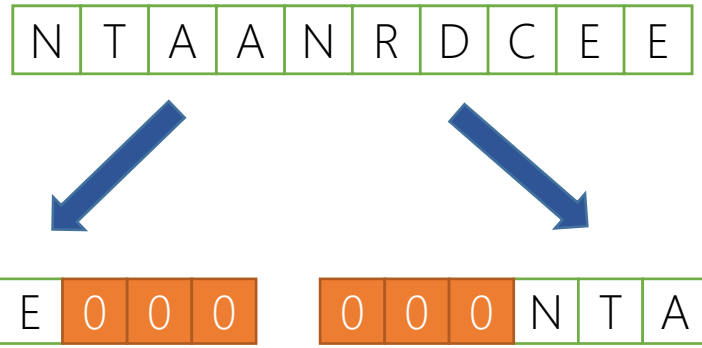


Panjang Awal = 10

gandakan sekuens awal
sebanyak maxpool size

Shift-and-Stitch CNN

CONV SIZE = 5
MAXPOOL SIZE = 2



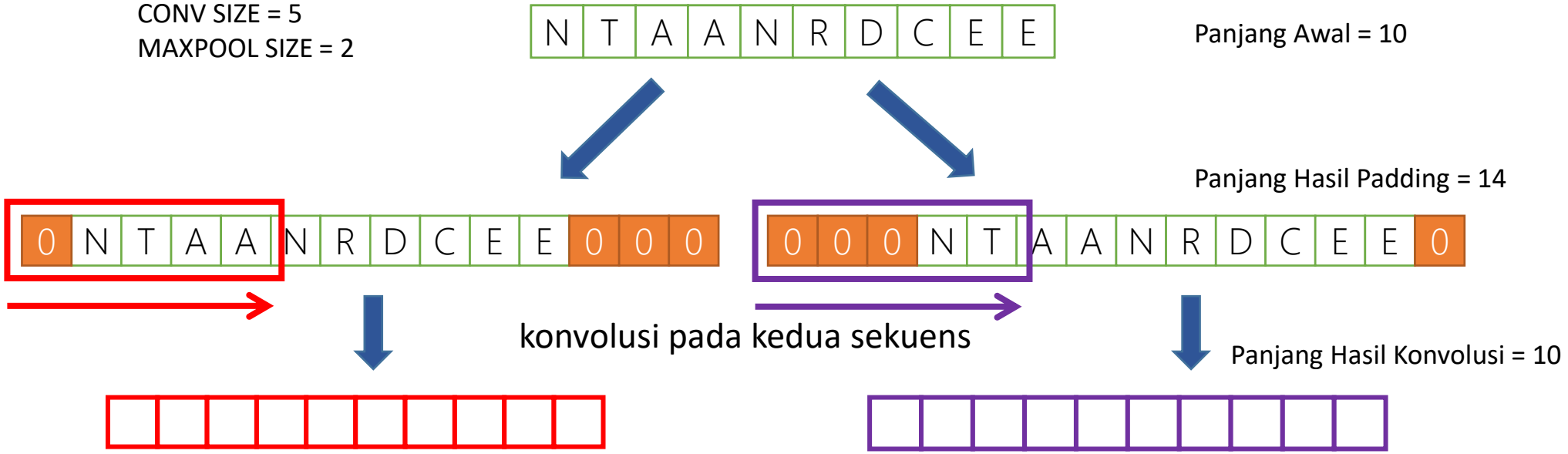
Panjang Awal = 10

Panjang Hasil Padding = 14

beri padding sebanyak:
(convsize - 1 / 2) - 1 di depan
(convsize - 1) - ((convsize - 1 / 2)- 1) di belakang

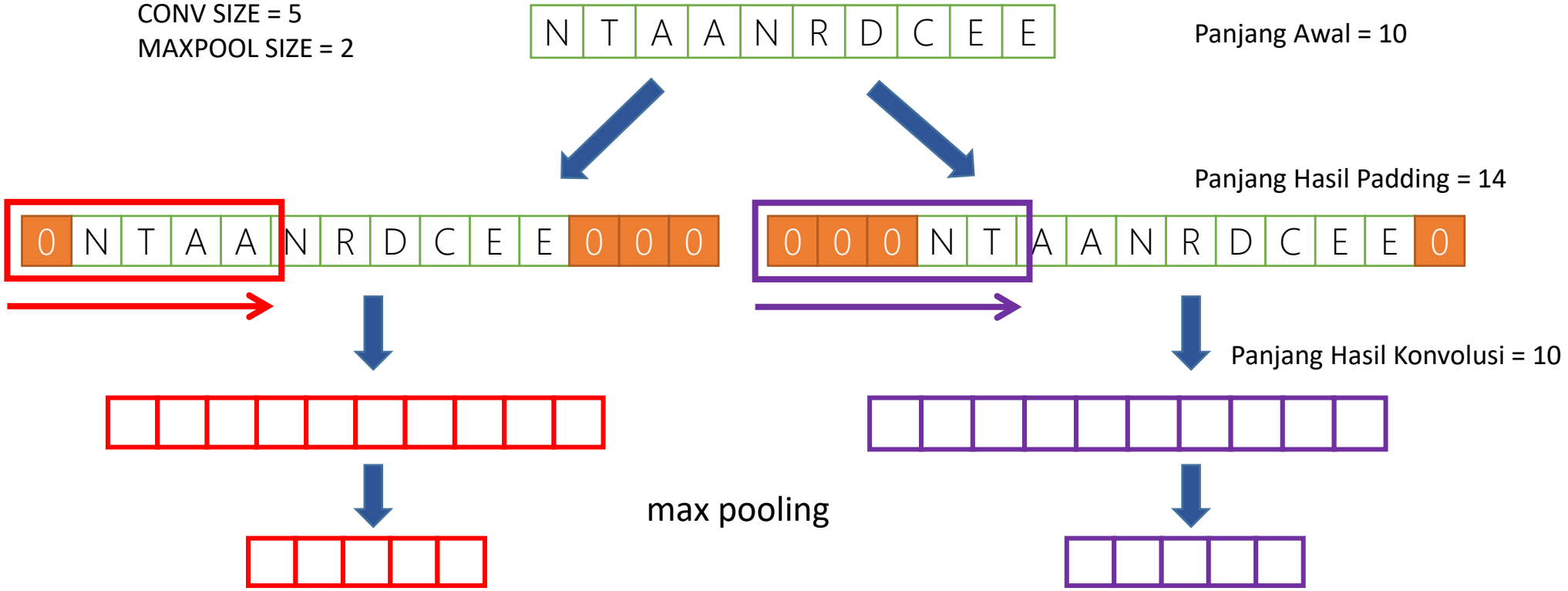
Shift-and-Stitch CNN

CONV SIZE = 5
MAXPOOL SIZE = 2



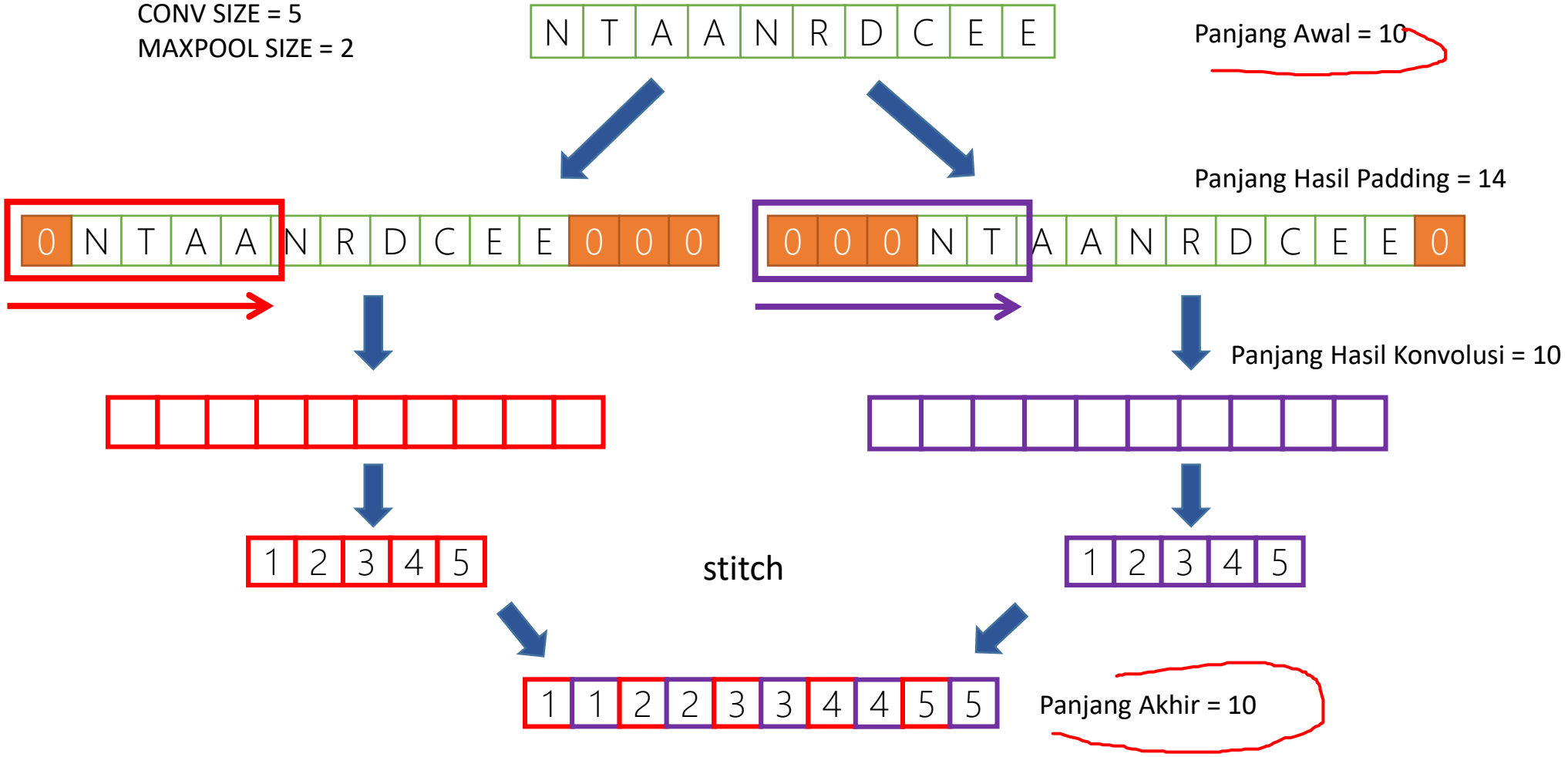
Shift-and-Stitch CNN

CONV SIZE = 5
MAXPOOL SIZE = 2

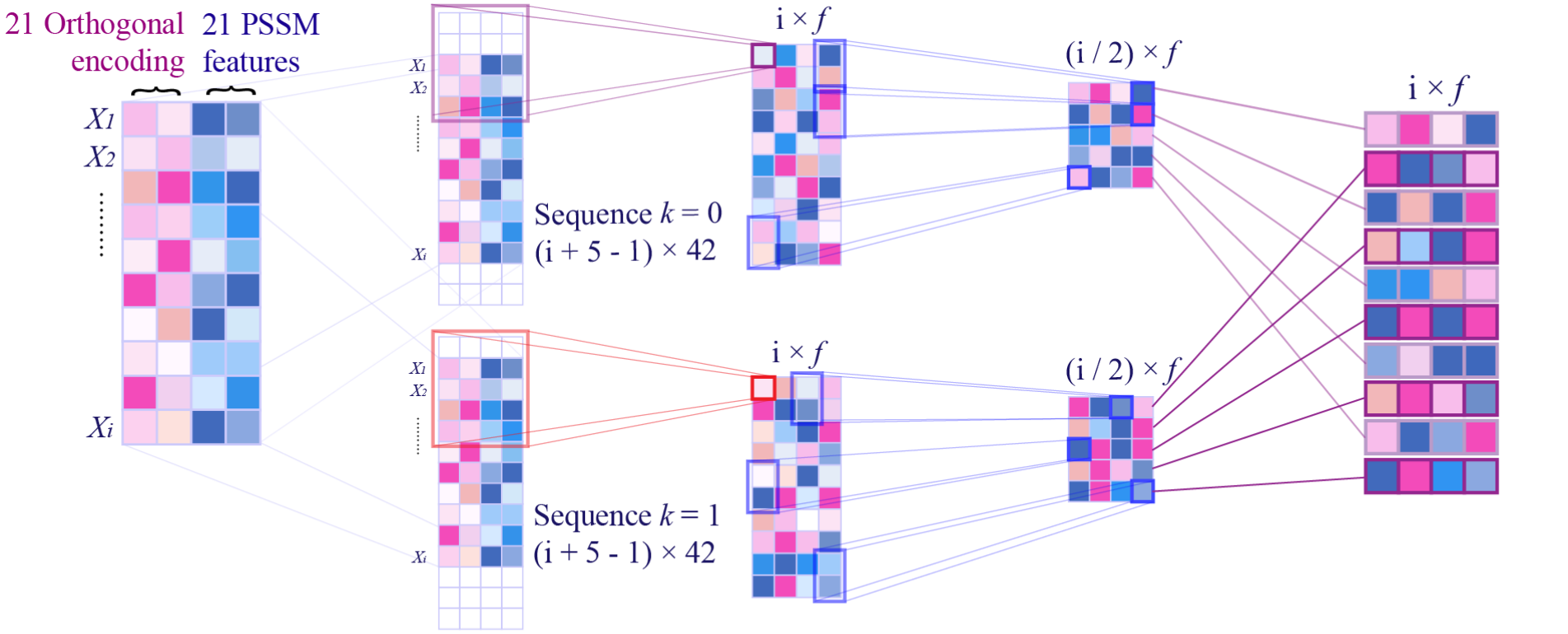


Shift-and-Stitch CNN

CONV SIZE = 5
MAXPOOL SIZE = 2



Shift-and-Stitch CNN



"Shift" on original sequence (i × 42)

convolution layers with convolution size $w = 5$

Max pooling layers

"Stitch" on shifted sequence

Shift-and-stitch sequence (i × f)

CNN Parameter Tuning

UNIVERSITAS GADJAH MADA

Nilai parameter yang tidak berubah

Parameter	Value
Dense Layers	1
Dense Layer Neurons	512
Input Dropout	0.2
Dropout	0.5
Activation Function	{ ReLU, Softmax }

Nilai parameter yang berubah

Parameter	Value
Convolution Layers	{ 2, 3, 4}
Convolution Size	{ 6, 7, 8, 9, 10 }
Feature maps	{ 512, 1024}

CNN Parameter Tuning

UNIVERSITAS GADJAH MADA

Nilai parameter yang tidak berubah

Parameter	Value
Dense Layers	1
Dense Layer Neurons	512
Input Dropout	0.2
Dropout	0.5
Activation Function	{ ReLU, Softmax }

Nilai parameter yang berubah

Parameter	Value
Convolution Layers	{ 2, 3, 4}
Convolution Size	{ 6, 7, 8, 9, 10 }
Feature maps	{ 512, 1024}

Convolutional Neural Network. Pada penelitian ini kombinasi arsitektur Convolutional Neural Network dan Support Vector Machine mampu menghasilkan tingkat akurasi yang lebih tinggi dibandingkan dengan arsitektur Convolutional Neural Network sederhana. Arsitektur kombinasi menghasilkan tingkat akurasi data uji CullPDB Q3 sebesar 80.75% dan Q8 sebesar 68.84% (Q3 meningkat 1.4% dan Q8 meningkat 0.71%), sementara itu pada data uji CB513 menghasilkan tingkat akurasi Q3 sebesar 78.51% dan Q8 sebesar 64.76% (Q3 meningkat 1.37% dan Q8 meningkat 0.67%).

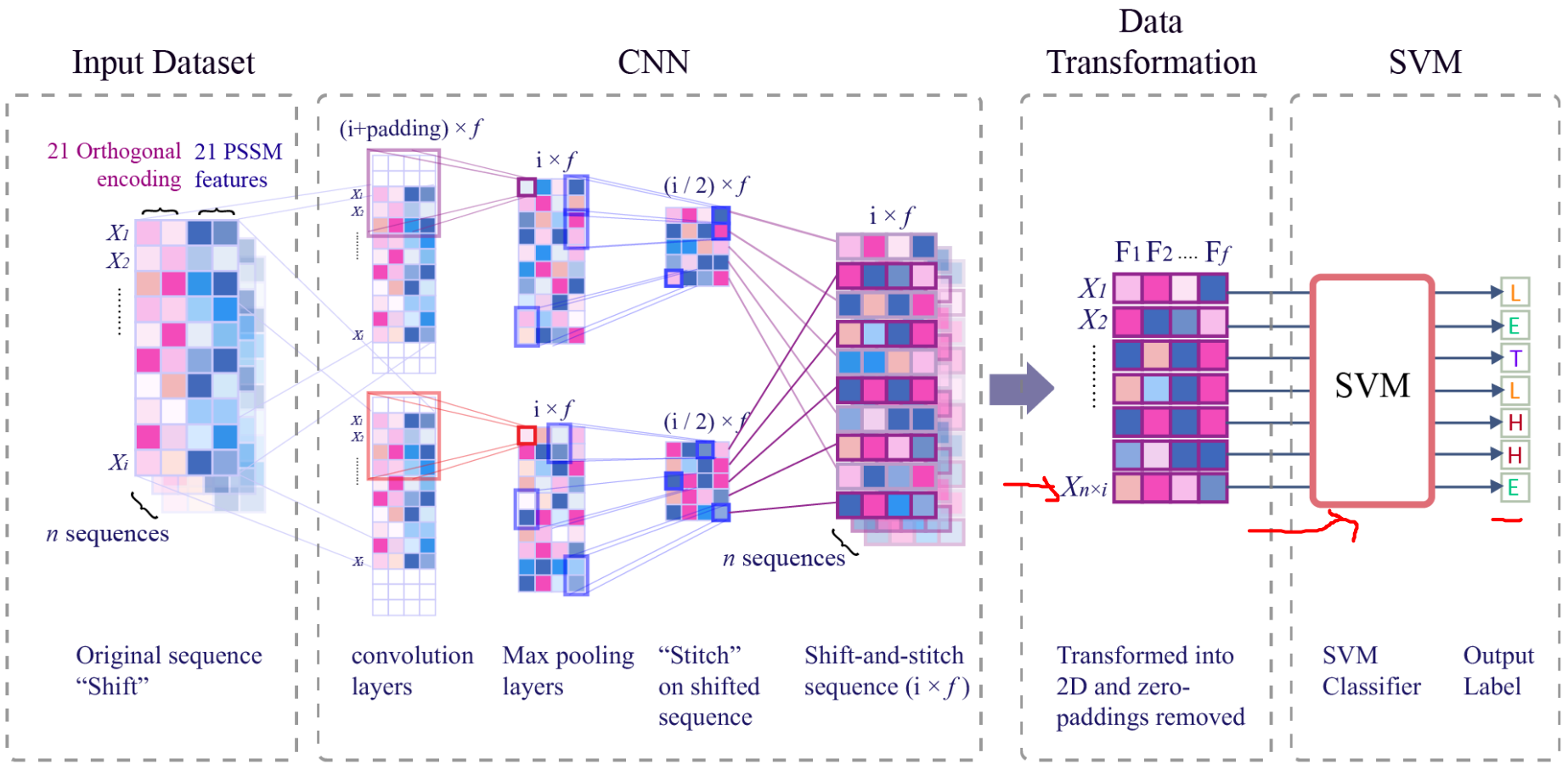
http://etd.repository.ugm.ac.id/penelitian/detail/186888

CNN Parameter Tuning

Convolution Layers	Convolution Size	Feature map	Accuracy (%)
2	6	512	67.701
2	7	512	67.509
2	8	512	67.990
2	9	512	68.012
2	10	512	68.097
3	6	512	68.521
3	7	512	68.329
3	8	512	68.387
3	9	512	68.442
3	10	512	68.291
4	6	512	68.467
4	**7**	**512**	**68.711**
4	8	512	68.535
4	9	512	68.594
4	10	512	68.344

Convolution Layers	Convolution Size	Feature map	Accuracy (%)
2	6	1024	67.918
2	7	1024	67.799
2	8	1024	67.731
2	9	1024	67.890
2	10	1024	67.750
3	6	1024	68.445
3	7	1024	68.266
3	8	1024	68.450
3	9	1024	68.468
3	10	1024	68.341
4	6	1024	68.486
4	7	1024	68.492
4	8	1024	68.434
4	9	1024	68.223
4	10	1024	68.174

CNN-SVM


Input Dataset
CNN
Data Transformation
SVM
21 Orthogonal encoding
21 PSSM features
X1
X2
Xi
n sequences
(i+padding) × f
i × f
(i / 2) × f
i × f
i × f
(i / 2) × f
n sequences
F1 F2 .... Ff
X1
X2
Xn×i
SVM
L
E
T
L
H
H
E
Original sequence "Shift"
convolution layers
Max pooling layers
"Stitch" on shifted sequence
Shift-and-stitch sequence (i × f)
Transformed into 2D and zero-paddings removed
SVM Classifier
Output Label

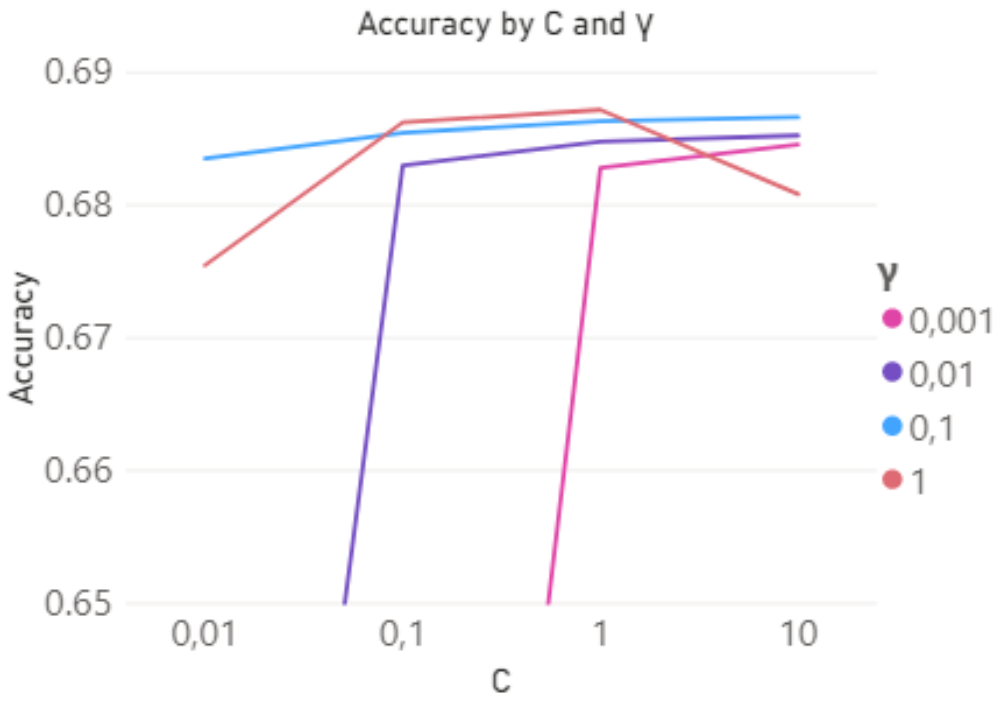
SVM Parameter Tuning

Nilai parameter yang berubah

Parameter	Value
C	{ 0.01, 0.1, 1, 10 }
γ	{ 0.001, 0.01, 0.1, 1

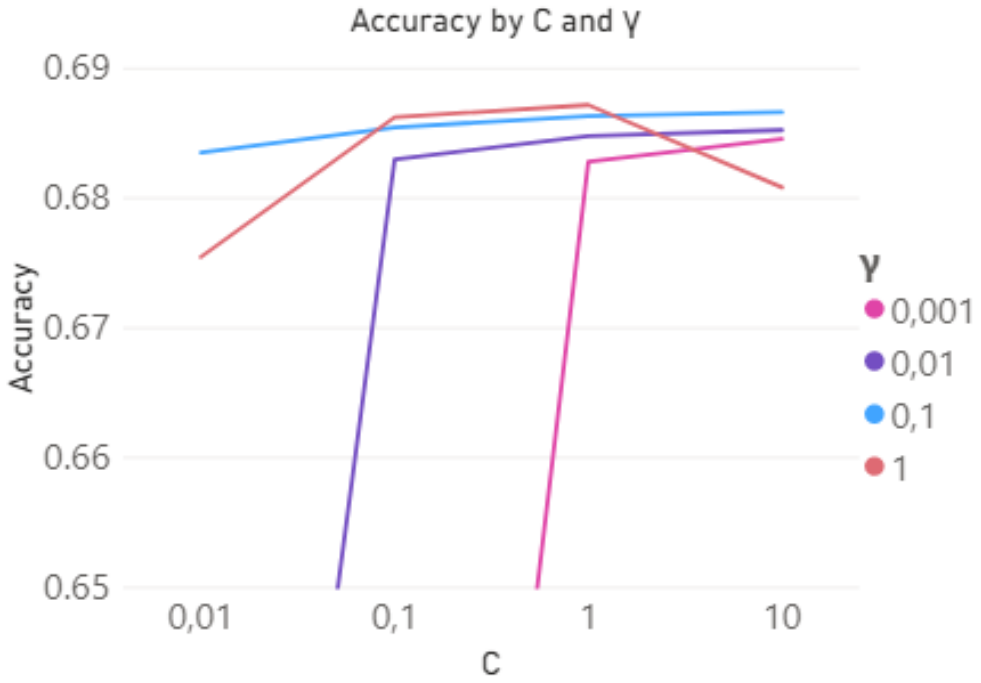
SVM Parameter Tuning

Parameters		Accuracy (%)
C	γ	
0.01	0.001	0.035
0.01	0.01	57.048
0.01	0.1	68.353
0.01	1	67.549
0.1	0.001	55.826
0.1	0.01	68.299
0.1	0.1	68.548
0.1	1	68.626
1	0.001	68.284
1	0.01	68.48
1	0.1	68.636
1	0.5	68.689
1	0.75	68.682
1	1	68.721
1	**1.25**	**68.735**
1	1.5	68.734
10	0.001	68.46
10	0.01	68.527
10	0.1	68.665
10	1	68.085



SVM Parameter Tuning

Parameters		Accuracy (%)
C	γ	
0.01	0.001	0.035
0.01	0.01	57.048
0.01	0.1	68.353
0.01	1	67.549
0.1	0.001	55.826
0.1	0.01	68.299
0.1	0.1	68.548
0.1	1	68.626
1	0.001	68.284
1	0.01	68.48
1	0.1	68.636
1	0.5	68.689
1	0.75	68.682
1	1	68.721
1	**1.25**	**68.735**
1	1.5	68.734
10	0.001	68.46
10	0.01	68.527
10	0.1	68.665
10	1	68.085



Perbandingan Performa

UNIVERSITAS GADJAH MADA

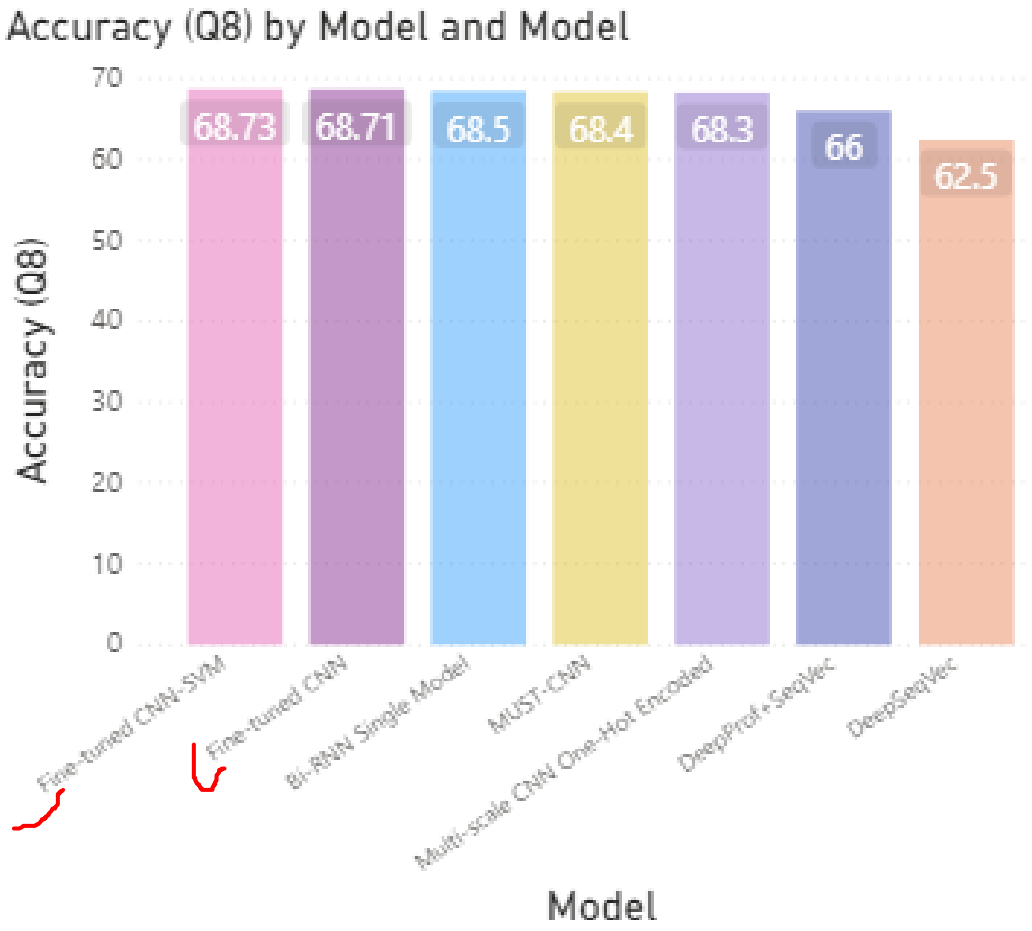
CNN (68.711%)

Label	Total Label	True Positive	Precision	Recall
H	26,157	24,045	0.8378	0.9193
E	18,016	14,525	0.7574	0.8062
L	17,920	11,576	0.5527	0.646
T	10,013	5,325	0.53	0.5318
S	8,316	1,909	0.4792	0.2296
G	3,132	834	0.454	0.2663
B	1,181	29	0.3867	0.0246
I	30	0	0	0
Total	**84,765**	**58,243**		

Hybrid CNN+SVM (68.735%)

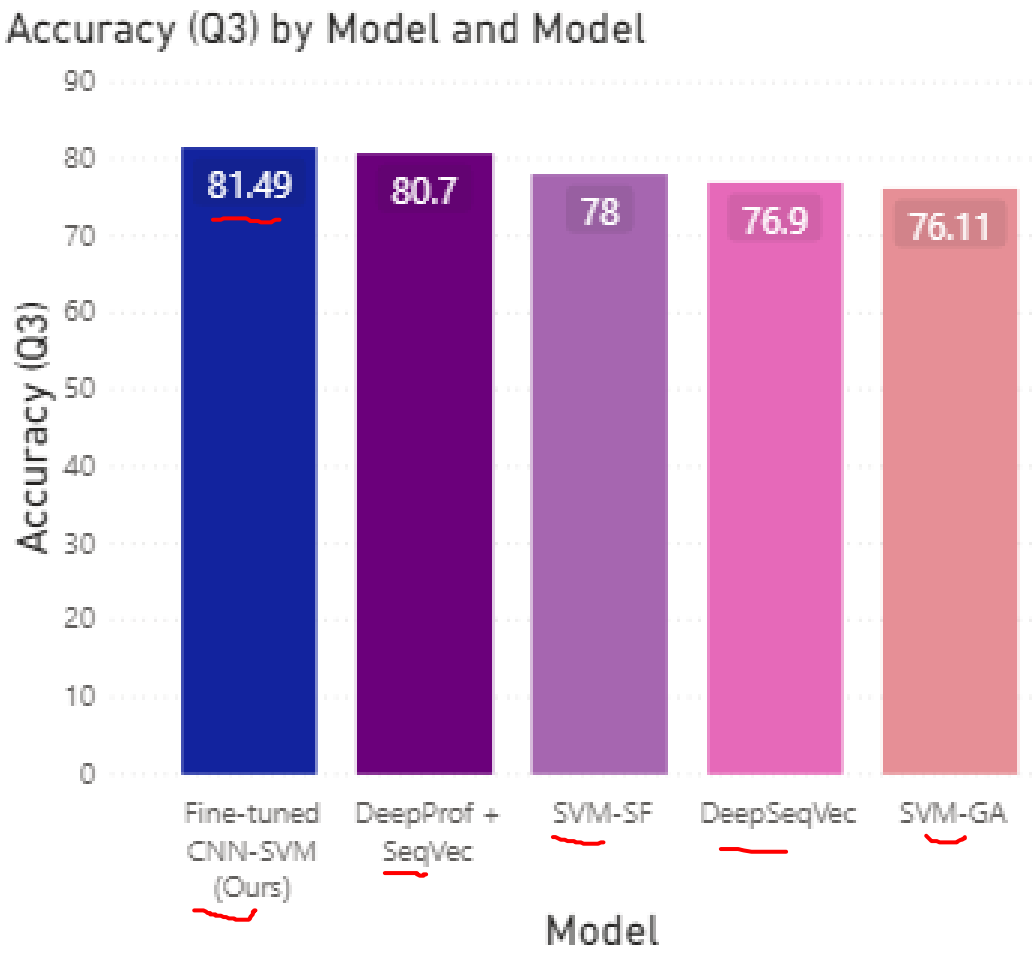
Label	Total Label	True Positive	Precision	Recall
H	26,157	23,739	0.8534	0.9076
E	18,016	14,412	0.7629	0.8
L	17,920	11,845	0.5480	0.661
T	10,013	5,339	0.5291	0.5332
S	8,316	1,986	0.4907	0.2388
G	3,132	896	0.4075	0.2861
B	1,181	46	0.4259	0.039
I	30	0	0	0
Total	**84,765**	**58,263**		

Perbandingan dengan Penelitian Lain



- Bi-RNN Single Model (Johansen et al., 2017)
- MUST-CNN (Lin et al., 2016)
- Multi-scale CNN One-hot Encoded (Zhou et al., 2018)
- DeepProf+SeqVec, DeepSeqVec (Heinzinger et al., 2019)

Perbandingan dengan Penelitian Lain



- DeepProf+SeqVec, DeepSeqVec (Heinzinger et al., 2019)
- SVM-SF (Wang et al., 2016)
- SVM-GA (Chen et al., 2016)

Kesimpulan

- Hybrid CNN-SVM untuk Sequence Labelling, spesifiknya untuk memprediksi struktur sekunder protein
- Hybrid CNN-SVM mampu meningkatkan kemampuan prediksi CNN biasa
- Menghasilkan skor Q3 81.49% dan Q8 68.735% pada data test CB513

UNIVERSITAS
GADJAH MADA

Predicting Secondary Structure of Protein Using Hybrid of Convolutional Neural Network and Support Vector Machine

International Journal of Intelligent Engineering & Systems
Vol.14, No.1, 2021

Terima Kasih!