Perbandingan Algoritma Klasifikasi dalam Memprediksi Mutasi Gen pada Data Sekuensing Menggunakan Metode Encoding

REZA ARDIANSYAH
232420019

PROGRAM STUDI TEKNIK INFORMATIKA S2
PROGRAM PASCASARJANA
UNIVERSITAS BINA DARMA
TAHUN 2025

PERBANDINGAN ALGORITMA KLASIFIKASI DALAM MEMPREDIKSI MUTASI GEN PADA DATA SEKUENSING MENGGUNAKAN METODE ENCODING

Tesis ini diajukan sebagai salah satu syarat
untuk memperoleh gelar

MAGISTER KOMPUTER

REZA ARDIANSYAH
ENTERPRISE IT INFRASTRUCTURE
232420019

PROGRAM STUDI TEKNIK INFORMATIKA – S2
PROGRAM PASCASARJANA
UNIVERSITAS BINA DARMA
PALEMBANG
2025

Halaman Pengesahan Pembimbing Tesis

Judul Tesis: **PERBANDINGAN ALGORITMA KLASIFIKASI DALAM MEMPREDIKSI MUTASI GEN PADA DATA SEKUENSING MENGGUNAKAN METODE ENCODING**

Oleh REZA ARDIANSYAH, NIM 232420019, Tesis ini telah disetujui dan disahkan oleh Pembimbing Program Studi Teknik Informatika – S2 konsentrasi ENTERPRISE IT INFRASTRUCTURE, Program Pascasarjana Universitas Bina Darma pada 10 September 2025 dan telah dinyatakan LULUS.

Palembang, 10 September 2025
Mengetahui,
Program Studi Teknik Informatika – S2
Universitas Bina Darma
Ketua,

Universitas Bina Darma
Magister Teknik Informatika

Dr. Usman Ependi, S.Kom., M.Kom.

Pembimbing,

Dr. Ahmad Nadar Mirzah, S.T., M.Kom.

Halaman Pengesahan Penguji Tesis

Judul Tesis: **PERBANDINGAN ALGORITMA KLASIFIKASI DALAM MEMPREDIKSI MUTASI GEN PADA DATA SEKUENSING MENGGUNAKAN METODE ENCODING**

Oleh REZA ARDIANSYAH , NIM 232420019, Tesis ini telah disetujui dan disahkan oleh Tim Penguji Program Studi Teknik Informatika – S2 konsentrasi ENTERPRISE IT INFRASTRUCTURE, Program Pascasarjana Universitas Bina Darma pada 10 September 2025 dan telah dinyatakan LULUS.

Palembang, 10 September 2025
Mengetahui,
Program Pascasarjana
Universitas Bina Darma
Direktur,

Prof. Dr. Ir. Achmad Syarifudin, M.Sc.

Penguji I,

Dr. Ahmad Haidar Mirzah, S.T., M.Kom.

Penguji II,

Prof. Dr. Edi Surya Negara, M.Kom.

Penguji III,

Dr. Tata Sutabri S.Kom, MMSI.

SURAT PERNYATAAN

Saya yang bertanda tangan dibawah ini:

Nama : REZA ARDIANSYAH

NIM : 232420019

Dengan ini menyatakan bahwa:

1. Karya tulis Saya (Tesis, Skripsi, Tugas Akhir) ini adalah asli dan belum pernah diajukan untuk mendapatkan gelar akademik baik (Magister, Sarjana, dan Ahli Madya) di Universitas Bina Darma;
2. Karya tulis ini murni gagasan, rumusan dan penelitian Saya sendiri dengan arahan tim pembimbing;
3. Dalam karya tulis ini tidak terdapat karya atau pendapat yang telah ditulis atau dipublikasikan orang lain, kecuali secara tertulis dengan jelas dikutip dengan mencantumkan nama pengarang dan memasukkan ke dalam daftar pustaka;
4. Karena yakin dengan keasliaan karya tulis ini, Saya menyatakan bersedia Tesis/Skripsi/Tugas Akhir, yang Saya hasilkan di unggah ke internet;
5. Surat Pernyataan ini Saya tulis dengan sungguh-sungguh dan apabila terdapat penyimpangan atau ketidakbenaran dalam pernyataan ini, maka Saya bersedia menerima sanksi dengan aturan yang berlaku di perguruan tinggi ini.

Demikian Surat Pernyataan ini saya buat agar dapat dipergunakan sebagaimana mestinya.

Palembang, 10 September 2025
Yang Membuat Pernyataan,

REZA ARDIANSYAH
NIM: 232420019

ABSTRAK

Klasifikasi sekuens merupakan tugas krusial dalam bioinformatika dan biologi komputasional, yang mendasari aplikasi seperti prediksi fungsi protein, klasifikasi penyakit, dan anotasi gen. Meskipun teknik *deep learning* telah berhasil di domain ini, kinerja model sangat bergantung pada cara sekuens dikodekan sebelum dimasukkan ke jaringan saraf tiruan. Studi yang ada telah mengeksplorasi berbagai skema pengkodean seperti One-Hot Encoding, k-mer, embedding, dan Position-Specific Scoring Matrics (PSSM), tetapi evaluasi komparatif yang komprehensif antar model masih terbatas. Studi ini bertujuan untuk mengatasi kesenjangan ini dengan mengevaluasi secara sistematis dampak empat strategi pengkodean terhadap kinerja dua model *deep learning* yang banyak digunakan: Convolutional Neural Networks (CNN) dan jaringan Bidirectional Long Short-Term Memory (BiLSTM). Dataset yang digunakan terdiri dari sekuens biologis berlabel (misalnya, protein atau DNA), yang telah diproses sebelumnya dan ditransformasikan menggunakan setiap metode pengkodean untuk evaluasi yang adil. Setiap model dilatih dan diuji dalam pengaturan yang konsisten untuk memastikan perbandingan yang andal. Hasil eksperimen menunjukkan bahwa pengkodean k-mer mengungguli semua metode lain, mencapai akurasi tertinggi dengan CNN (90%) dan BiLSTM (90%). Representasi berbasis embedding juga memberikan hasil yang kuat, terutama dengan BiLSTM, yang memanfaatkan kemampuannya untuk menangkap dependensi jarak jauh dalam data sekuensial, untuk One-Hot Encoding juga menunjukan hasil yang baik. Sebaliknya, PSSM menunjukkan akurasi yang jauh lebih rendah di kedua model, menunjukkan keterbatasan daya representasionalnya untuk tugas *deep learning*. Tujuan penelitian ini adalah untuk memandu praktisi dalam memilih strategi pengkodean yang optimal untuk model *deep learning* dalam tugas klasifikasi sekuens. Dengan mengidentifikasi kombinasi model-pengodean yang paling efektif, studi ini berkontribusi pada peningkatan akurasi prediktif dan efisiensi komputasi dalam aplikasi bioinformatika.

Kata Kunci: Deep-Learning,CNN, BiLSTM, Sequence Classification, K-mer

ABSTRACT

Sequence classification is a critical task in bioinformatics and computational biology, underpinning applications such as protein function prediction, disease classification, and gene annotation. Despite the success of deep learning techniques in this domain, model performance is highly dependent on the way sequences are encoded before being input to neural networks. Existing studies have explored various encoding schemes such as One-Hot Encoding, k-mer, embeddings, and Position-Specific Scoring Matrices (PSSM), but a comprehensive comparative evaluation across different models remains limited. This study aims to address this gap by systematically evaluating the impact of four encoding strategies on the performance of two widely used deep learning models: Convolutional Neural Networks (CNN) and Long Short-Term Memory (LSTM) networks. The dataset used comprises labeled biological sequences (e.g., protein or DNA), pre-processed and transformed using each encoding method for fair evaluation. Each model was trained and tested under consistent settings to ensure reliable comparison. Experimental results reveal that k-mer encoding outperforms all other methods, achieving the highest accuracy with both CNN (90%) and LSTM (90%). Embedding-based representations also deliver strong results, particularly with LSTM, which leverages its ability to capture long-range dependencies in sequential data, for One-Hot Encoding also shows good results. In contrast, PSSM demonstrated significantly lower accuracy across both models, suggesting limitations in their representational power for deep learning tasks. The goal of this research is to guide practitioners in selecting optimal encoding strategies for deep learning models in sequence classification tasks. By identifying the most effective model-encoding combinations, this study contributes to improving predictive accuracy and computational efficiency in bioinformatics applications.

Keywords: *Deep-Learning,CNN, BiLSTM, Sequence Classification, k-mer*

MOTTO DAN PERSEMBAHAN

MOTTO

"Kita Mungkin Bisa Menunda,
Tapi Waktu Tidak Akan Menunggu"

"Tuhan Bersumpah Demi Waktu, Manusia Duji Dengan Cara Ia Memanfaatkannya, Setiap Detik Yang Disia-Siakan Bukan Hanya Waktu Yang Hilang, Tetapi Penyesalan Yang Menanti Diakhir Perjalanan."

PERSEMBAHAN

Kupersembahkan Thesis ini untuk:

- Yang Utama Dari Segalanya Puji syukur kepada Allah SWT. Taburan cinta dan kasih sayang-Mu telah memberikan kekuatan, membekaliku dengan ilmu serta memperkenalkanku dengan cinta. Atas karunia serta kemudahan yang Engkau berikan akhirnya *thesis* ini dapat terselesaikan. Sholawat dan salam selalu terlimpahkan keharibaan Rasullah Muhammad SAW.
- Kedua orang tuaku ayah Sutriawan dan ibu Anita Fitriana yang telah berjuang keras, dan selalu mendoakan ku, membimbing, memberikan nasehat semangat dan ikhtiar ,tanpa kalian berdua aku bukanlah siapa siapa dan tanpa kalian aku tidak akan bisa mencapai keberhasilan. setiap Doa dan tetesan keringatmu adalah motivasi keberhasilan bagi hidupku ini.

- Kakek dan nenek yang selalu mendokanku,memberikan dukungan dalam menggapai cita cita.
- Terimakasih kakakku Armila ferolita yang selalu mendokan keberhasilanku, membantu memberi semangat untukku.yang sangat berarti bagi pembutan *thesis* ini, dan juga untuk hidup ku sekarang semoga ini tidak akan pernah berakhir.
- Para sahabat serta Teman-teman seperjuanganku terutama program studi magister teknik informatika, terimkasih atas pertemuan ,dukungan dan keakraban selama ini
- Almamater tercintaku program studi magister teknik informatika Fakultas ilmu komputer Universitas Bina Darma Palembang.

KATA PENGANTAR

Puji syukur kepada Allah SWT berkat Rahmat, Hidayah, dan Karunia-Nya kepada kita semua sehingga penulis dapat menyelesaikan *thesis* ini dengan judul Perbandingan Algoritma Klasifikasi dalam Memprediksi Mutasi Gen pada Data Sekuensing Menggunakan Metode Encoding**.** Penulis menyadari dalam penyusunan *thesis* ini tidak akan selesai tanpa bantuan dari berbagai pihak. Karena itu pada kesempatan ini penulis ingin mengucapkan terima kasih kepada :

1. Prof. Dr. Sunda Ariana, M.Pd., M.M. Selaku Rektor Universitas Bina Darma Palembang
2. Prof. Dr. Ir. H. Achmad Syarifudin, M.Sc.. Selaku direktur pascasarjana Universitas Bina Darma
3. Dr. Usman Ependi, M.Kom. Selaku Kepala Program Studi Teknik Informatika Fakultas Pascasarjana Universitas Bina Darma
4. Dr. Ahmad Haidar Mirzah S.T., M.Kom. Selaku Dosen pembimbing saya yang telah membantu dalam penyusunan tesis ini
5. Prof. Dr. Edi Surya Negara, M.Kom. Selaku Dosen Penguji satu saya
6. Dr. Tata Sutabri, S.Kom, MMSI. Selaku Dosen Penguji dua saya
7. Bapak/Ibu dosen dan staf/karyawan di Fakultas Pascasarjana yang telah berkontribusi dalam kelancaran penulisan tesis ini.
8. Bapak dan Ibu serta keluarga besar saya yang tentunya berperan sangat penting dalam kelancaran penyusunan tesis ini berkat dukungan motivasi dan bantuan dari mereka penulis lebih semangat untuk menyelesaikan tesis ini

9. Sahabat dan teman-teman seperjuangan angkatan MTI 29 A yaitu Majduddin, Egy Septian, Aria Dinata, Nyimas Hamidah PA, Titah, Nur Ayu Wulandari dan Annisa Fitri Aulia.

Penulis juga menyadari sepenuhnya bahwa di dalam *thesis* ini terdapat kekurangan dan jauh dari kata sempurna. Oleh sebab itu, kami berharap adanya kritik, saran dan usulan demi perbaikan *thesis* yang telah kami buat di masa yang akan datang, mengingat tidak ada sesuatu yang sempurna tanpa saran yang membangun.

Diharapkan, *thesis* ini bisa bermanfaat untuk semua pihak. Selain itu, kritik dan saran yang membangun sangat penulis harapkan dari pembaca sekalian agar *thesis* ini bisa lebih baik lagi.

Palembang, September 2025

Reza Ardiiansyah

DAFTAR ISI

Halaman

DAFTAR TABEL

Halaman

DAFTAR GAMBAR

Halaman

DAFTAR LAMPIRAN