KLASIFIKASI GEN ESENSIAL PADA *DROSOPHILA MELANOGASTER* BERDASARKAN PROTEIN *SEQUENCE* MENGGUNAKAN METODE *LONG SHORT-TERM MEMORY* (LSTM)

(Skripsi)

Oleh

DINA PUTRI AULIA

1957051004

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM

UNIVERSITAS LAMPUNG

BANDAR LAMPUNG

2023

ABSTRAK

KLASIFIKASI GEN ESENSIAL PADA *DROSOPHILA MELANOGASTER* BERDASARKAN PROTEIN *SEQUENCE* MENGGUNAKAN METODE *LONG SHORT-TERM MEMORY* (LSTM)

Oleh

Dina Putri Aulia

Gen esensial adalah gen yang sangat diperlukan untuk mendukung kehidupan organisme seluler. Penghapusan gen esensial pada organisme akan mengakibatan kematian dan infertilitas. Mengklasifikasikan gen esensial ada 2 cara, yaitu dengan teknik eksperimental dan komputasi, namun teknik eksperimental membutuhkan sumber daya yang besar. Pada penelitian ini mengusulkan metode *Long Short-Term Memory* (LSTM) untuk mengklasifikasikan gen esensial pada *Drosophila melanogaster* berdasarkan protein *sequence*. *Drosophila melanogaster* adalah salah satu organisme yang sering digunakan untuk analisis *science*. Tujuan penelitian ini, yaitu mengukur dan mengklasifikasi hasil penelitian menggunakan algoritma *Long Short-Term Memory* (LSTM) dan membandingkan hasilnya dengan penelitian terdahulu oleh Beder, et al, (2021). Data dibagi dengan 2 skenario, yaitu : 80% *training* 20% *validation* ; dan 90% *training* 10% data *validation*. Pada dataset CEG memiliki distribusi kelas yang tidak seimbang sehingga dilakukan proses *Random Undersampling* (RUS) untuk menyeimbangkan kelas. Evaluasi model dilakukan dengan metrik evaluasi, yaitu PR-AUC, ROC-AUC, *sensitivity* dan *specificity*. Hasil kinerja yang paling baik pada dataset OEG didapatkan pada skenario pembagian data 80% *training* dan 20% *validation*, dengan nilai *sensitivity* 81%, *specificity* 76%, ROC-AUC 79% , PR-AUC 82%. Hasil yang paling baik pada dataset CEG diperoleh dari pembagian data 80% *training* dan data 20% *validation* dengan nilai yang didapat untuk *sensitivity* 73%, *specitifity* 50% ROC-AUC 61%, dan PR-AUC 45%.

Kata Kunci : Gen Esensial, Klasifikasi, *Drosophila melanogaster*, Protein *sequence*, LSTM.

ABSTRACT

CLASSIFICATION OF ESSENTIAL GENES IN DROSOPHILA MELANOGASTER BASED ON PROTEIN SEQUENCE USING LONG SHORT-TERM MEMORY UNIT (LSTM)

Oleh

Dina Putri Aulia

Essential genes are genes that are very necessary to support the life of cellular organisms. Deletion of essential genes in an organism will result in death and infertility. There are 2 ways to classify essential genes, specifically experimental and computational techniques, but experimental techniques require large resources. In this study, we propose the Long Short-Term Memory (LSTM) method to classify essential genes in Drosophila melanogaster based on protein sequences. Drosophila melanogaster is an organism that is often used for scientific analysis. The purpose of this research is to measure and classify research results using the Long Short-Term Memory (LSTM) algorithm and compare the results with previous research by Beder, et al, (2021). The data is divided into 2 scenarios, specifically 80% training 20% validation and 90% training 10% data validation. The CEG dataset has an unbalanced class distribution so a Random Undersampling (RUS) process is carried out to balance the classes. Model evaluation was carried out using evaluation metrics, specifically PR-AUC, ROC-AUC, sensitivity and specificity. The best performance results on the OEG dataset were obtained in the data sharing scenario of 80% training and 20% validation, with sensitivity values of 81%, specificity 76%, ROC-AUC 79%, PR-AUC 82%. The best results on the CEG dataset were obtained from dividing 80% training data and 20% validation data with values obtained for sensitivity 73%, specificity 50% ROC-AUC 61%, and PR-AUC 45%

Keyword : Gen Essential, Classification, Drosophila melanogaster, Protein sequence, LSTM.

KLASIFIKASI GEN ESENSIAL PADA *DROSOPHILA MELANOGASTER* BERDASARKAN PROTEIN *SEQUENCE* MENGGUNAKAN METODE *LONG SHORT-TERM MEMORY* (LSTM)

Oleh

DINA PUTRI AULIA

Skripsi

Sebagai Salah Satu Syarat untuk Mencapai Gelar
SARJANA ILMU KOMPUTER

Pada

Jurusan Ilmu Komputer
Fakultas Matematika dan Ilmu Pengetahuan Alam
Universitas Lampung

FAKULTAS MATEMATIKA DAN ILMU PENGETAHUAN ALAM
UNIVERSITAS LAMPUNG
BANDAR LAMPUNG
2023

Judul Skripsi	: **KLASIFIKASI GEN ESENSIAL PADA *DROSOPHILA MELANOGASTER* BERDASARKAN PROTEIN *SEQUENCE* MENGGUNAKAN METODE *LONG SHORT-TERM MEMORY (LSTM)***
Nama Mahasiswa	: Dina Putri Aulia
Nomor Pokok Mahasiswa	: **1957051004**
Program Studi	: Ilmu Komputer
Fakultas	: Matematika dan Ilmu Pengetahuan Alam

MENYETUJUI

1. Komisi Pembimbing

Favorisen R. Lumbanraja, Ph.D.
NIP 19830110 200812 1 002

2. Ketua Jurusan Ilmu Komputer

Didik Kurniawan, S.Si., M.T.
NIP 19800419 200501 1 004

MENGESAHKAN

1. Tim Penguji

   Ketua : **Favorisen R. Lumbanraja, Ph.D.** .....

   Penguji I
   Penguji Pembahas : **Fatma Indriani, S.T., M.I.T., Ph.D.** .....

   Penguji II
   Penguji Pembahas : **Dr. rer. nat Akmal Junaidi, M.Sc.** .....

2. Dekan Fakultas Matematika dan Ilmu Pengetahuan Alam

Dr. Eng. Heri Satria, S.Si., M.Si.
NIP 19711001 200501 1 002

Tanggal Lulus Ujian Skripsi : **22 Desember 2023**

PERNYATAAN

Saya yang bertanda tangan di bawah ini :

Nama : Dina Putri Aulia

NPM : 1957051004

Dengan ini menyatakan bahwa skripsi saya yang berjudul "KLASIFIKASI GEN ESENSIAL PADA *DROSOPHILA MELANOGASTER* BERDASARKAN PROTEIN *SEQUENCE* MENGGUNAKAN METODE *LONG SHORT-TERM MEMORY* (LSTM)" adalah benar hasil karya sendiri dan bukan orang lain. Seluruh tulisan yang tertuang dalam skripsi ini telah mengikuti kaidah penulisan karya ilmiah Universitas Lampung. Jika di kemudian hari terbukti skripsi saya adalah hasil penjiblakan atau dibuat oleh orang lain, maka saya bersedia menerima sanksi sesuai dengan peraturan yang berlaku.

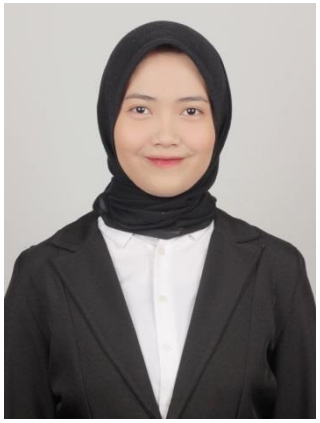
Bandar Lampung, 23 Januari 2024

Penulis

Dina Putri Aulia

NPM 1957051004

RIWAYAT HIDUP



Penulis dilahirkan di Gading Rejo pada tanggal 12 Agustus 2002 sebagai anak ketiga dari tiga bersaudara dari pasangan Bapak Kamarudin dan Ibu Nazaria. Penulis menyelesaikan Pendidikan Sekolah Dasar (SD) di SD Negeri 7 Gading Rejo pada tahun 2014. Kemudian melanjutkan Pendidikan Sekolah Pertama (SMP) di SMP Negeri 1 Gading Rejo yang diselesaikan pada tahun 2017. Kemudian melanjutkan Pendidikan Sekolah Menengah Atas (SMA) di SMA Negeri 1 Gading Rejo yang diselesaikan pada tahun 2019.

Penulis terdaftar sebagai mahasiswa Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung pada tahun 2019 melalui jalur Seleksi Mandiri Masuk Perguruan Tinggi Negeri (SMMPTN) Wilayah Barat. Selama menjadi mahasiswa, penulis melakukan beberapa kegiatan antara lain.

1. Menjadi anggota Adapter Himpunan Mahasiswa Jurusan Ilmu Komputer pada periode 2019/2020.
2. Menjadi anggota pengurus dibidang Eksternal Himpunan Mahasiswa Jurusan Ilmu Komputer (Himakom) pada periode 2018/2019.
3. Menjadi Asisten Dosen Jurusan Ilmu Komputer untuk mata kuliah Multimedia pada periode semester ganjil tahun ajaran 2021/2022.
4. Menjadi Asisten Dosen Jurusan Biologi Terapan untuk mata kuliah DasarDasar Bioinformatika pada periode semester genap tahun ajaran 2022/2023.

5. Melaksanakan Kerja Praktik di Pringsewu pada tahun 2022.
6. Melaksanakan Kuliah Kerja Nyata (KKN) pada tahun ajaran 2022/2023 di Desa Banding Agung, Talang Padang, Tanggamus, Lampung.

MOTTO

لَا يُكَلِّفُ اللهُ نَفْسًا اِلَّا وُسْعَهَا

“Allah tidak membebani seseorang melainkan sesuai dengan kesanggupannya”

(Q.S Al-Baqarah, 2 : 286)

إِنَّ مَعَ الْعُسْرِ يُسْرًا ﴿ ٦﴾

فَإِذَا فَرَغْتَ فَانْصَبْ ﴿ ٧﴾ وَإِلَىٰ رَبِّكَ فَارْغَبْ ﴿ ٨﴾

“Sesungguhnya sesudah kesulitan itu ada kemudahan, Maka apabila kamu telah selesai (dari sesuatu urusan), kerjakanlah dengan sungguh-sungguh (urusan) yang lain, dan hanya kepada Tuhanmulah hendaknya kamu berharap.”

(Q.S Al-Insyirah, 94 : 6-8)

“ Long story short, i survived”

(Taylor Swift)

PERSEMBAHAN

Alhamdulillahirobbilalamin

Puji syukur kepada Allah SWT yang telah memberikan rahmat dan hidayah-Nya sehingga dapat menyelesaikan penulisan skripsi ini. Sholawat dan salam saya sanjungkan kepada Nabi Muhammad SAW.

Aku persembahkan karya ini kepada :

Ayah dan Mama

Terima kasihku ucapkan kepada Ayah dan Mama yang tercinta. Terima kasih atas segala pengorbanan dan tulus kasih yang diberikan. Terima kasih telah mendidik dan membesarkanku dengan kasih sayang kalian. Terima kasih karena selalu melangitkan doa doa baik untuk kesuksesanku. Terima kasih atas semua perjuangan kalian yang tiada hentinya. Terima kasih Ayah dan Mama.

Kakakku Ayu Dita Destiani dan Nadia Komala Dewi

Terima kasih atas segala doa dan dukungan yang telah diberikan kepada saya selama ini.

Seluruh Keluarga Besar, Sahabat, dan Teman-teman yang selalu memberikan semangat dan dukungan.

Almamater Tercinta, Universitas Lampung

SANWACANA

Puji syukur kehadirat Allah SWT, karena telah memberikan rahmat dan hidayah-Nya kepada saya sehingga saya dapat menyelesaikan skripsi dengan judul "Klasifikasi Gen Esensial Pada *Drosophila Melanogaster* Berdasarkan Protein *Sequence* Menggunakan Metode *Long Short-Term Memory*". Saya berharap skripsi ini dapat menambah pengetahuan bagi pembaca tentang gen esensial, protein *sequence*, *Drosophila melanogaster* dan metode *long short-term memory*.

Selama proses penulisan skripsi ini tidak terlepas dari dukungan banyak pihak yang telah membimbing, membantu dan memberikan dukungan kepada saya, sehingga pada kesempatan ini saya ingin menyampaikan ungkapan terima kasih kepada:

1. Kedua orang tua serta kedua saudara kandung yang paling berjasa dalam hidup saya. Terimakasih atas kepercayaan yang selalu diberikan kepada saya dan seluruh do'a serta dukungan yang tiada henti oleh mama dan ayah selama saya hidup.
2. Bapak Favorisen R. Lumbanraja, Ph. D. sebagai pembimbing utama yang telah membimbing saya dengan memberikan kritik dan saran serta membina dalam menyelesaikan skripsi ini yang dapat diselesaikan dengan baik.
3. Ibu Fatma Indriani, S.T., M.I.T., Ph.D. sebagai pembahas utama yang telah membimbing saya dengan memberikan ide, kritik dan saran serta membina dalam menyelesaikan skripsi ini yang dapat diselesaikan dengan baik.

4. Bapak Dr. rer. nat. Akmal Junaidi, M. Sc. sebagai pembahas kedua yang telah memberikan ide, kritik, dan saran serta membina dalam menyelesaikan skripsi ini yang dapat diselesaikan dengan baik.
5. Bapak Rizky Prabowo, M.Kom. selaku dosen pembimbing akademik, yang telah membimbing selama perkuliahan di Jurusan Ilmu Komputer.
6. Bapak Dr. Eng. Heri Satria, S.Si., M.Si selaku Dekan FMIPA Universitas Lampung.
7. Bapak Didik Kurniawan, S.Si., M.T., selaku Ketua Jurusan Ilmu Komputer Universitas Lampung.
8. Ibu Anie Rose Irawati, S.T., M.Cs. selaku sekretaris Jurusan Ilmu Komputer FMIPA Universitas Lampung.
9. Bapak dan Ibu Dosen Jurusan Ilmu Komputer FMIPA Universitas Lampung yang telah memberikan ilmu dan pengalaman dalam hidup untuk menjadi lebih baik.
10. Ibu Ade Nora Maela, Bang Zainuddin, dan Mas Nofal yang telah membantu segala urusan administrasi dan segala jenis izin penulis di Jurusan Ilmu Komputer.
11. Zameena Kaureen Almahyra, keponakan tersayang yang selalu menjadi penyemangat dan mood boster bagi penulis.
12. Azahra Alya Hidayah, Jihan Cahya Fatimah, Ardella Dean Awalia, dan Mohammad Fajar sebagai teman seperbimbingan yang senantiasa membantu, menyemangati, dan menguatkan satu sama lain
13. Eldi Jauhari Saputra, Takhfa Nur Asyifa, Azahra Alya Hidayah, Anastasya Dian Nuratri, dan Lulu Vania Nariswari sebagai seseorang yang selalu ada ketika senang dan sedih. Terimakasih karena selalu mendengarkan keluh kesah penulis, memberikan dukungan, perhatian, dan do’a sampai saat ini.
14. Seluruh teman satu Kuliah Kerja Nyata (KKN) Rara Gusti Rahmawati, Daffara Rifqia Putri, Bernika Febriyanti, Royyan Fajrul Falah, dan Alvian Firmansyah yang selalu menghibur dan memberikan semangat kepada penulis dalam proses pengerjaan skripsi.
15. Teman-teman Jurusan Ilmu Komputer FMIPA Universitas Lampung angkatan 2019 yang telah memberikan cerita dalam masa perkuliahan.

16. Semua pihak yang telah berpartisipasi baik secara langsung maupun tidak langsung dalam membantu penyusunan skripsi ini.

Penulis menyadari bahwa dalam penulisan skripsi ini masih terdapat banyak kekurangan karena keterbatasan kemampuan, pengalaman serta pengetahuan penulis. Oleh karena itu, saran dan kritik yang membangun sangat diharapkan sebagai bahan evaluasi untuk kedepannya. Semoga skripsi ini dapat bermanfaat bagi semua pihak.

Bandar Lampung, 23 Januari 2024

Dina Putri Aulia
NPM. 1957051004

DAFTAR ISI

Halaman

DAFTAR TABEL

Tabel Halaman

DAFTAR GAMBAR

Gambar Halaman

DAFTAR KODE PROGRAM

Kode Program Halaman

I. PENDAHULUAN

1.1.Latar Belakang

Gen esensial adalah gen yang sangat diperlukan untuk mendukung kehidupan organisme seluler (Zhang, et al., 2004). Penghapusan gen esensial pada organisme akan mengakibatan kematian dan infertilitas. Gen esensial diklasifikasikan sebagai gen yang penting untuk kelangsungan hidup atau reproduksi suatu organisme pada keadaan tertentu. Gen esensial memiliki peran penting dalam perkembangan suatu organisme pada kondisi umum dan juga sebagai gen penting yang menopang fungsi seluler utama (Peng, et al., 2017). Untuk bertahan hidup, suatu organisme harus dapat melakukan setidaknya dua fungsi mendasar, yaitu memperoleh energi dan bereproduksi. Sel harus mengubah nutrisi dari makanan, seperti protein, lemak dan gula, menjadi adenosin trifosfat (ATP), sumber energi utama bagi sel hidup. Proses konversi energi ini disebut sebagai metabolisme (Zhang & Ren, 2015).

Mengetahui gen esensial untuk kelangsungan hidup organisme merupakan hal yang sangat penting untuk pemahaman tentang mekanisme dasar kehidupan dan kebutuhan hidup pada organisme seluler. Identifikasi gen esensial memberikan informasi mengenai esensialitas yang digunakan dalam berbagai penelitian, misalnya untuk menemukan penyakit gen manusia (Steinmetz, et al., 2002), pemahaman dalam pembuatan obat, terapi dalam penyakit kanker (Sharma, Eils & Konig, 2016), atau mengidentifikasi target. insektisida untuk membunuh serangga yang merugikan dalam pertanian. Secara fungsional gen esensial berperan

dalam proses pemeliharaan sel-sel yang mendasar seperti sintesis protein, DNA, dan RNA.

Dalam mengidentifikasi gen esensial, lalat buah (*Drosophila melanogaster*) adalah serangga dalam kategori ordo diptera yang menjadi organisme model yang paling banyak digunakan dalam penelitian genetika berbagai proses perkembangan dan pewarisan (Miklos & Rubin, 1996). *Drosophila melanogaster* adalah salah satu organisme pertama yang digunakan untuk analisis protein dalam memahami perilaku protein pada eukariota lain, termasuk manusia (Cerniker & Rubin, 2003). Jadi, penting untuk mempelajari fungsionalitas protein dalam organisme ini karena mengarah ke lebih banyak informasi dalam jaringan manusia. Identifikasi gen esensial menjadi salah satu masalah penting dalam genomic komputasi. Ada dua pendekatan yang digunakan untuk menentukan gen esensial, yaitu metode eksperimental dan komputasi (Liu, et al., 2017). Namun, metode eksperimen memakan waktu yang lama serta biaya yang mahal dan apabila memakai metode eksperimen yang berbeda dapat menghasilkan hasil yang berbeda (Xu, et al., 2011). Maka dari itu, metode prediksi komputasi menawarkan alternatif yang baik dengan menggunakan berbagai pendekatan ekstraksi fitur untuk klasifikasi dan pelatihan.

Metode *machine learning* untuk memprediksi gen esensial telah banyak digunakan sebagai akumulasi pengurutan data untuk organisme jumlah besar serta kumpulan gen dan protein esensial. Seperti halnya penelitian yang dilakukan oleh Aromolaran, et al. (2020) untuk memprediksi gen esensial pada *Drosophila melanogaster* menggunakan metode *Generalized Linear Model* (GLM), *Support Vector Machine* (SVM), *Random Forest* (RF), *Artificial Neural Network* (NNET), dan *Extreme Gradient Boosting* (XGB). Selain itu, Khanh Le, et al. (2020) dengan memanfaatkan model pemrosesan bahasa alami dan *machine learning* dalam mempelajari *sequence* biologis menggunakan metode *k-nearest neighbors* (kNN), *Random Forest* (RF), *Support Vector Machine* (SVM), *multi-layer*

perceptron (MLP), *convolutional neural network* (CNN). Selain metode *machine learning* yang digunakan untuk klasifikasi protein *sequence*, metode *deep learning* juga dipakai untuk mengklasifikasi pola kompleks data protein dalam skala besar. Beberapa penelitian juga telah menggunakan metode *deep learning* dalam mengklasifikasikan protein yang mengarah pada penemuan ilmiah.

Pada penelitian ini mengusulkan metode *Long Short-Term Memory* (LSTM) untuk mengklasifikasikan gen esensial pada *Drosophila melanogaster* berdasarkan protein *sequence*. *Long Short-Term Memory* adalah sistem *recurrent neural* yang kuat dan dirancang khusus untuk mengatasi masalah yang biasanya muncul saat mempelajari depedensi jangka panjang (Houdt & Nápoles, 2020). Penelitian ini dilakukan untuk mengukur dan mengklasifikasi hasil penelitian menggunakan algoritma *Long Short-Term Memory* (LSTM).

1.2. Rumusan Masalah

Berdasarkan pemaparan dari latar belakang yang dibuat, maka rumusan masalah pada penelitian ini adalah sebagai berikut :

1. Apakah metode *Long Short-Term Memory* (LSTM) dapat diimplementasikan untuk membuat model klasifikasi pada protein *sequence Drosophila melanogaster*?
2. Berapa hasil evaluasi kinerja yang didapatkan dari metode *Long Short-Term Memory* (LSTM) dalam mengklasifikasikan protein *sequences* pada *Drosophila melanogaster*?
3. Apakah hasil yang diperoleh lebih baik dari penelitian terdahulu oleh Beder, et al. (2021)?

1.3.Batasan Masalah

Batasan masalah dalam penelitian ini, yaitu :

1. Proses klasifikasi gen esensial didasarkan pada protein *sequence* dan menggunakan arsitektur model *Long Short-Term Memory* (LSTM).
2. Klasifikasi dilakukan dengan 2 kelas, yaitu esensial dan *non*-esensial.
3. Data yang digunakan berupa data *Cellular* Esensial *Gene* (CEG) dan *Organismal* Esensial *Gene* (OEG) dari organisme *Drosophila melanogaster* yang diperoleh dari penelitian Beder, et al. (2021).

1.4.Tujuan Penetilian

Tujuan dari penelitian ini adalah sebagai berikut :

1. Mengukur dan mengklasifikasi hasil penelitian menggunakan algoritma *Long Short-Term Memory* (LSTM).
2. Membandingkan hasil yang diperoleh terhadap penelitian terdahulu yang menggunakan dataset yang sama dengan metode klasifikasi yang digunakan.

1.5.Manfaat Penelitian

Manfaat dari penelitian ini adalah sebagai berikut :

1. Menambah pengetahuan dan wawasan tentang cara kerja algoritma *Long Short-Term Memory* (LSTM) dalam melakukan klasifikasi gen esensial pada *Drosophila Melanogaster* dengan protein *sequence.*
2. Dapat mengetahui kinerja dari metode *Long Short-Term Memory* (LSTM) dalam mengklasifikasi gen esensial pada *Drosophila Melanogaster*.

II. TINJAUAN PUSTAKA

2.1.Penelitian Terdahulu

Penelitian ini dilakukan tidak lepas dari penelitian sebelumnya, sehingga penelitian yang dilakukan memiliki hubungan antara persamaan dan perbedaan objek dengan penelitian sebelumnya. Ringkasan dari penelitian terdahulu dapat dilihat pada Tabel 1.

Tabel 1. Penelitian Terdahulu Terkait Klasifikasi Gen Esensial

No	Penelitian	Data	Metode	Hasil
1	*A Computational Framework Based on Ensemble Deep Neural Networks for* Esensial *Genes Identification* (Khanh Le, et al., 2020)	DEG (*Database of* Esensial *Genes*) Esensial *Gene* : 518 *Non*-esensial : 1.072	*k-nearest neighbors* (kNN), *Random Forest* (RF), *Support Vector Machine* (SVM), *Multi-Layer Perceptron* (MLP), *Convolutional Neural Network* (CNN).	Akurasi pada metode : kNN : 73,2% RF : 73,6% SVM : 74% MLP : 74,8% CNN : 74,7%

2	Esensial *gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features* (Aromolaran, et al., 2020)	Database : OGEE (*Online GENe* Esensial*ity*) dan DEG (*Database of Gen* Esensial) Esensial *Gene* : 441 *Non*-esensial : 11.788	*Generalized Linear Model* (GLM), *Support Vector Machine* (SVM), *Random Forest* (RF), *Artificial Neural Network* (NNET), dan *Extreme Gradient Boosting* (XGB)	Testing Set **XGB :** ROC-AUC : 0,90 PR-AUC : 0,30 F1 : 0,34 **GLM :** ROC-AUC : 0,89 PR-AUC : 0,27 F1 : 0,28 **SVM :** ROC-AUC : 0,88 PR-AUC : 0,27 F1 : 0,30 **NNET :** ROC-AUC : 0,85 PR-AUC : 0,20 F1 : 0,24 RF : ROC-AUC : 0,90 PR-AUC : 0,29 F1 : 0,32

3	*Identifying* esensial *genes across eukaryotes by machine learning* (Beder, et al., 2021)	Database : OGEE (*Online GENe* Esensial*ity*) dan DEG (*Database of Gen* Esensial) Esensial *Gene* : 11.038 *Non*-esensial : 67.035	*Randon Forest* (RF) *dan Extreme Gradient Boosting* (XGB)	Testing Set **RF :** ROC-AUC : 0,86 PR-AUC : 0,60 *Sensitivity* : 0,65 *Specificity* : 0,85 **XGB :** ROC-AUC : 0,85 PR-AUC : 0,59 *Sensitivity* : 0,67 *Specificity* : 0,83

Penelitian terdahulu digunakan sebagai acuan dalam penelitian ini. Penjelasan mengenai penelitian terdahulu yang digunakan adalah sebagai berikut :

2.1.1. *A Computational Framework Based on Ensemble Deep Neural Networks for Essential Genes Identification*

Penelitian ini dilakukan oleh Khanh Le, et al. (2020). Penelitian ini memanfaatkan model pemrosesan bahasa alami dalam mempelajari *sequence* biologis. Dalam penelitian ini, pendekatan berbasis pembelajaran mesin dibangun sebagai upaya untuk memberikan wawasan komprehensif tentang patologi, kehidupan, dan evolusi. Metode yang digunakan, yaitu *k-nearest neighbors* (kNN), *Random Forest* (RF), *Support Vector Machine* (SVM), *multi-layer perceptron* (MLP), *convolutional neural network* (CNN). Penelitian ini menggunakan data yang diambil dari *Database of* Esensial *Genes* (DEG) dengan jumlah 518 gen esensial, dan 1072

gen *non*-esensial. Metode CNN dan MLP berkinerja terbaik di antara keempatnya, yaitu dengan hasil akurasi 74,8%, koefisien korelasi Matthews (MCC) 0,385, nilai AUC 0,775 untuk metode MLP dan akurasi 74,7% koefisien korelasi Matthews (MCC) 0,384, nilai AUC 0,775 untuk metode CNN dan diikuti oleh model SVM.

2.1.2. *Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features*

Penelitian ini dilakukan oleh Aromolaran, et al. (2020). Pada penelitian ini memprediksi gen esensial pada *Drosophila melanogaster* dengan metode *machine learning* berdasarkan pada berbagai macam aspek berbeda yang terdiri dari urutan nukleotida dan protein, jaringan gen, interaksi protein, konservasi evolusioner, dan anotasi fungsional. Metode *machine learning* yang digunakan, yaitu metode *Generalized Linear Model* (GLM), *Support Vector Machine* (SVM), *Random Forest* (RF), *Artificial Neural Network* (NNET), dan *Extreme Gradient Boosting* (XGB). Dengan menggunakan data daftar gen esensial yang dikumpulkan dari basis data *Online GENe* Esensial*ity* (OGEE) dan *Database of Gen* Esensial (DEG). Data yang diambil sebanyak 441 gen esensial dan 11.788 gen *non*-esensial.

Hasil dari penelitian ini menunjukkan bahwa metode *Extreme Gradient Boosting* (XGB) berkinerja terbaik dengan menghasilkan ROC AUC = 0,90, PR-AUC = 0,30 dan F1 = 0,34. PR-AUC dan F1 mengukur kinerja prediksi positif terhadap total pengamatan positif, dimana semakin tinggi skor semakin baik model yang digunakan, terutama ketika memprediksi kelas positif pada fokus analisis seperti yang ada dalam penelitian ini.

2.1.3. *Identifying Essential Genes Across Eukaryotes by Machine Learning*

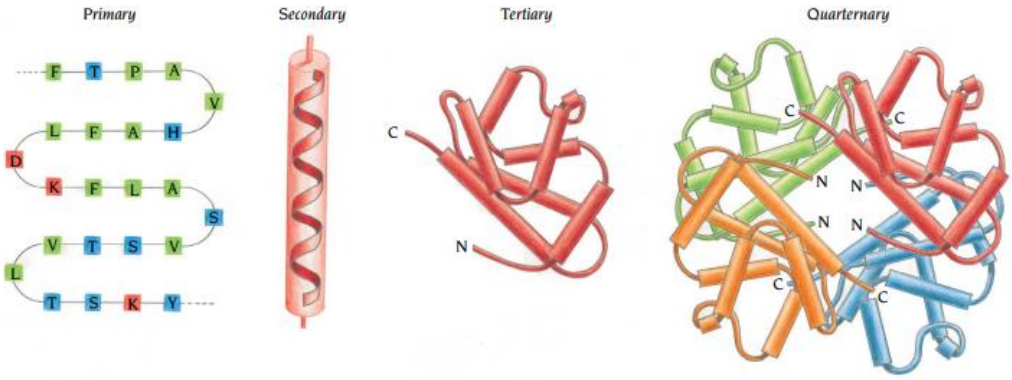
Penelitian ini dilakukan oleh Beder, et al. (2021). Penelitian ini dilakukan dengan tujuan untuk mengidentifikasi gen esensial pada eukariota menggunakan 6 organisme, yaitu *C.elegan, D.melanogaster, H.sapiens, M.otot, S.cerevisiae dan S.pombe* yang didapat kan dari database *Online GENe* Esensial*ity* (OGEE) dan *Database of Gen* Esensial (DEG). Pengklasifikasi dilatih berdasarkan data yang dari lima organisme dan divalidasi dengan organisme keenam. Data yang didapatkan sebanyak 11.038 gen esensial dan 67.35 gen non-esensial. Pada penelitian ini metode *machine learning* yang digunakan, yaitu metode *Random Forest* (RF) *dan Extreme Gradient Boosting* (XGB). Berdasarkan hasil dari *testing sets* metode *Random Forest* menghasilkan nilai ROC-AUC sebesar 0,86, PR-AUC sebesar 0,60, *Sensitivity* sebesar 0,65 dan *Specificity* sebesar 0,85. Sedangkan, metode *Extreme Gradient Boosting* menghasilkan nilai ROC-AUC sebesar 0,85, PR-AUC sebesar 0,59, *Sensitivity* sebesar 0,67 dan *Specificity* sebesar 0,83.

2.2.Protein

Protein merupakan untaian sederetan residu asam amino dengan urutan spesifik yang dihubungkan oleh ikatan peptida (Azhar, 2016). Protein tersusun dari peptida-peptida sehingga membentuk suatu polimer yang disebut polipeptida. Setiap monomernya tersusun atas suatu asam amino. Protein menyediakan tidak hanya karbon dan hidrogen tetapi juga nitrogen dan sulfur yang tidak tersedia pada lemak dan karbohidrat. Beberapa fungsi protein, yaitu sebagai sumber energi, berperan besar dalam membangun dan memperbaiki jaringan tubuh, membentuk antibodi , dan sebagai katalis reaksi metabolik. Protein dapat didefinisikan sebagai senyawa organik kompleks dengan struktur dasar yang tersusun dari 20

jenis asam amino yang berbeda dan saling berikatan (Fairuz, et al., 2022).

Struktur asam amino secara umum, yaitu terdiri dari satu atom C yang mengikat empat gugus : gugus amina (NH_2), gugus karboksil (COOH), atom hidrogen (H), dan satu gugus sisa (R) atau bisa disebut juga rantai samping yang membedakan satu asam amino dengan asam yang lainnya (Suprayitno & Sulistiyati, 2017). Asam amino dapat diklasifikasikan esensial yang sangat penting untuk metabolisme protein dan *non*-esensial yang diperlukan untuk fungsi sel normal dan dapat disintesis dari asam amino lain dalam tubuh untuk memberikan protein jaringan agar tubuh dapat menggunakannya. Karena protein tersusun dari asam amino yang berbeda secara kimiawi, maka suatu protein akan terangkai melalui ikatan peptida yang terkadang dihubungkan oleh ikatan sulfida. Sehingga protein bisa mengalami pelipatan-pelipatan membentuk struktur yang bermacam-macam. Adapun gambaran dari struktur protein yang dapat dilihat pada Gambar 1.



Gambar 1. Tingkatan Struktur Protein (Branden & Tooze, 2012).

Menurut (Azhar, 2016) protein dikelompokkan menjadi 4 tingkatan struktur, yaitu :

1. Struktur Primer

   Urutan rangkaian residu asam amino yang dihasilkan dari pembentukan ikatan peptide antara gugus α–amino dengan gugus α–karboksil. Ikatan peptide yang merupakan ikatan kovalen adalah ikatan

yang memelihara struktur primer. Struktur ini dapat menentukan urutan suatu asam amino dari suatu polipeptida.

2. Struktur Sekunder

   Struktur terbentuk akibat ikatan hidrogen antara hidrogen amida dan oksigen karbonil pada rantai ikatan peptida dari protein tersebut. Struktur sekunder protein berbentuk *α-helix, β-sheet, loop*, dan *turn*. Struktur *α-helix* dan *β-sheet* merupakan struktur yang berulang pada interval yang teratur. Struktur *α-helix* seperti spiral dan hanya melibatkan satu polipeptida, sedangkan struktur *β-sheet* dapat melibatan satu atau lebih polipeptida.

3. Struktur Tersier

   struktur tersier protein berkaitan dengan lapisan selanjutnya dari struktur sekunder. Interaksi non-kovalen antara rantai samping residu asam amino dan ikatan kovalen disulfida memainkan peranan yang menentukan struktur tersier protein. Sehingga, penataan keseluruhan struktur tiga-dimensi dari semua atom-atom di dalam suatu protein direfer sebagai struktur tersier protein.

4. Struktur Quartener

   Penataan polipeptida pada protein multisubunit didalam struktur tiga-dimensinya dinamakan struktur quartener. Struktur quartener memiliki hubungan dengan topologi penataan ruang dari dua atau lebih rantai polipeptida. Struktur quartener merupakan penataan dan pengorganisasian subunit protein menjadi protein kompleks yang fungsional. Interaksi antara subunit pada protein multisubunit dimediasi oleh interaksi non-kovalen seperti ikatan hidrogen, interaksi hidrofobik, dan interaksi elektrostatik.

2.3. *Drosophila Melanogaster*

Drosophila melanogaster adalah jenis serangga yang termasuk dalam anggota ordo diptera atau bangsa lalat (McLaughlin & Bratu, 2015). *Drosophila melanogaster* telah diaplikasikan secara luas untuk menjelaskan berbagai fenomena biologis penting yang juga terdapat pada manusia, mulai dari peran apoptosis dan fagositosis dalam perkembangan dan imunitas (Nainu, 2018). *Drosophila melanogaster* merupakan hewan yang tidak bertulang belakang dengan ukuran tubuh sekitar 3 mm. *Drosophila melanogaster* dalam sistematika taksonomi, dapat diklasifikasikan sebagai berikut (O'Grady & Markow, 2009) :

Kingdom : Animalia
Phylum : Arthropoda
Subphylum : Hexapoda
Class : Insecta
Ordo : Diphtera
Family : Drosophilidae
Genus : Drosophila
Spesies : Drosophila melanogaster

Genom serangga ini berukuran sekitar 180 MB (megabasa) yang tersebar pada empat kromosom (Adams, et al., 2000). Dengan jumlah kromosom yang sedikit, *Drosophila melanogaster* kemudian menjadi organisme pilihan untuk mempelajari mekanisme penyusunan gen pada kromosom, pengaturan aktivitas dan fungsi gen, serta pola mutasi pada organisme eukariotik sederhana. Walaupun memiliki genom yang sederhana, *Drosophila melanogaster* diperkirakan memiliki kemiripan genetik dengan manusia sebesar 75% . Hal inilah yang mendasari potensi penggunaan lalat buah atau *Drosophila melanogaster* sebagai organisme model dalam riset mekanisme penyakit dan penemuan obat. Gambaran dari organisme *Drosophila melanogaster* dapat di lihat pada Gambar 2.

Gambar 2. *Drosophila Melanogaster* (Perveen, 2017).

2.4.Gen esensial

Gen esensial diklasifikasikan sebagai gen yang penting untuk kelangsungan hidup atau reproduksi suatu organisme pada keadaan tertentu. Gen esensial memiliki peran penting dalam perkembangan suatu organisme pada kondisi umum dan juga sebagai gen penting yang menopang fungsi seluler utama (Peng, et al., 2017). Untuk bertahan hidup, suatu organisme harus dapat melakukan setidaknya dua fungsi mendasar, yaitu memperoleh energi dan bereproduksi. Sel harus mengubah nutrisi dari makanan, seperti protein, lemak dan gula, menjadi adenosin trifosfat (ATP), sumber energi utama bagi sel hidup. Proses konversi energi ini disebut sebagai metabolisme (Zhang & Ren, 2015).

Dalam organisme, gen tidak berfungsi secara independen, interaksi antara gen atau protein ada dan menjaga stabilitas lingkungan internal. Jaringan biologis kompleks yang terdiri dari gen yang berinteraksi umumnya bebas skala, yang berarti bahwa hanya ada beberapa node yang sangat terhubung dan banyak node yang jarang terhubung dalam jaringan (Xingyi Li, et al., 2019). Gen esensial mengandung informasi kunci genom oleh sebab itu bisa menjadi kunci pemahaman komprehensif tentang kehidupan. Selain itu, karena gen esensial mempunyai peran penting dalam biologi sintetik,

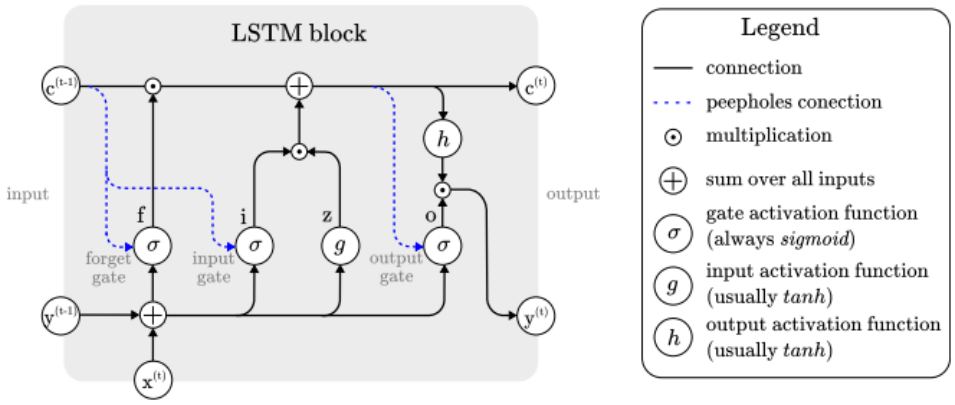
gen esensial sangat penting untuk pengembangan genom (Khanh Le, et al., 2020). Pemahaman komprehensif tentang gen esensial dapat memungkinkan para ilmuwan untuk menjelaskan sifat biologis mikroorganisme, menghasilkan subset gen minimal, dapat membantu mengembangkan target obat, dan menghasilkan obat potensial untuk mengobati penyakit menular.

2.5.*Preprocessing* Data

Preprocessing data adalah tahap pertama yang dilakukan dengan mengubah data mentah dari format yang tidak beraturan ke dalam bentuk yang mudah dipahami. Data *preprocessing* bertujuan untuk memperkecil ukuran data, menemukan relasi antar data, menormalkan data, menghilangkan *outlier*, dan mengekstraksi fitur untuk data yang termasuk beberapa teknik seperti pembersihan data, integrasi, transformasi dan pengurangan (Alasadi & Bhaya, 2017).

2.6.*Long Short-Term Memory* (LSTM)

Long Short-Term Memory adalah sistem *recurrent neural* yang kuat dan dirancang khusus untuk mengatasi masalah yang biasanya muncul saat mempelajari depedensi jangka panjang (Houdt, Mosquera & Nápoles, 2020). Arsitektur LSTM terdiri dari satu set sub-jaringan yang terhubung secara berulang, yang dikenal sebagai blok memori. Metode *Long Short-Term Memory* merupakan metode RNN tingkat lanjut yang mampu menangani masalah *gradient* yang hilang. Struktur dasar metode LSTM dapat dilihat pada Gambar 3 dan penjelasan mengenai notasi arsitektur metode *Long Short-Term Memory* (LSTM) dapat dilihat pada Tabel 2.



Gambar 3. Arsitektur Dasar Metode *Long Short-Term Memory* (LSTM) (Houdt, Mosquera, & Nápoles, 2020).

Tabel 2. Notasi Arsitektur Metode *Long Short-Term Memory* (LSTM) (Smagulova & James, 2020)

$\mathbf{x_t}$	***Input vector***
h_{t-1}	*Output of a previous cell*
C_t	*Cell memory of current state*
C_{t-1}	*Cell memory of a previous cell*
$\tilde{C}_t$	*Candidate to a cell memory*
i_t	*Input gate*
o_t	*Output gate*
f_t	*Forget gate*
g_t	*Input gate*
σ	*sigmoid function*
Tanh	*hyperbolic tangent function*
$W^{(*)}$, $U^{(*)}$, $V^{(*)}$	*weight matrices*
b^*	*Biases*

Jaringan *Long Short-Term Memory* terdiri dari blok memori berbeda yang disebut *cell*. *Cell* yang akan berikan kepada *cell* berikutnya disebut *cell state* dan *hidden state*. Blok *memory* bertanggung jawab untuk mengingat sesuatu dan manipulasi terhadap *memory* dilakukan melalui tiga mekanisme utama yang disebut *gates.* Berikut merupakan penjelasan dari struktur *Long Short-Term Memory* :

2.6.1. *Forget Gate*

Forget gate bertanggung jawab untuk menghapus informasi dari *cell state.* Suatu informasi yang tidak diperlukan atau kurang penting akan dihapus melalui penggandaan filter. Penghapusan informasi yang kurang penting perlu dilakukan untuk mengoptimalkan kinerja *Long Short-Term Memory*. *Forget gate* didefinisikan pada Persamaan (1) (Smagulova & James, 2020).

$$f_t = \sigma\ (w^{(f)}x_t\ +\ U^{(f)}h_{t-1} +\ b^{(f)})\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Forget gate mengambil dua input, yaitu h_{t-1} dan x_t, h_{t-1} adalah *hidden state* dari *cell* sebelumnya atau keluaran dari *cell* sebelumnya dan x_t adalah masukan pada langkah waktu tertentu. *Input* yang diberikan dikalikan dengan matriks bobot. Selanjutnya, fungsi sigmoid mengeluarkan vektor, dengan nilai mulai dari 0 hingga 1, sesuai dengan setiap angka dalam *cell state*. Pada dasarnya, fungsi sigmoid bertanggung jawab untuk memutuskan nilai mana yang akan dipertahankan dan mana yang akan dibuang. Jika 0 adalah *output* untuk nilai tertentu dalam *cell state*, maka informasi akan dihapus atau dilupakan. Demikian pula, jika nilainya 1, maka informasi akan diingat. Keluaran vektor dari fungsi sigmoid ini dikalikan dengan *cell state.*

2.6.2. *Input Gate*

Input gate bertanggung jawab untuk menambahkan informasi ke dalam *cell state* dan mengatur nilai apa saja yang perlu ditambahkan pada *cell state* dengan melibatkan fungsi sigmoid. *Input gate* membuat vektor yang berisi semua nilai yang mungkin dapat ditambahkan ke dalam *cell state*, dilakukan dengan menggunakan fungsi tanh, yang mengeluarkan nilai dari -1 ke +1. Selanjutnya, mengalikan nilai filter regulasi (gerbang sigmoid) ke vektor yang dibuat (fungsi tanh). *Input gate* didefinisikan pada Persamaan (2) (Smagulova & James, 2020).

$i_t = \sigma\ (w^{(i)}x_t + U^{(i)}h_{t-1} + b^{(i)})$……..……..……(2)

2.6.3. *Output Gate*

Fungsi *output gate*, yaitu membuat vektor setelah menerapkan fungsi tanh pada *cell state*, dengan meningkatkan nilai ke kisaran -1 hingga +1, membuat filter menggunakan nilai h_{t-1} dan x_t, sehingga dapat mengatur nilai-nilai yang perlu dari vektor yang dibuat menggunakan fungsi sigmoid. *Output gate* didefinisikan pada Persamaan (3) (Smagulova and James, 2020).

$o_t = \sigma\ (w^{(o)}x_t + U^{(o)}h_{t-1} + b^{(o)})$……...……...……(3)

contoh implementasi perhitungan *Forward propagation* pada *Long Short-Term Memory* sebagai berikut.

Data *input* pada *timestep* t_0 dan t_1 , yaitu :

$$x_0 = \begin{bmatrix} 0.25 \\ 0.30 \end{bmatrix} \quad \text{dengan label} \quad \begin{bmatrix} 0.5 \\ 0.3 \end{bmatrix}$$

$$x_1 = \begin{bmatrix} 1 \\ 0.80 \end{bmatrix} \quad \text{dengan label} \quad \begin{bmatrix} 1.25 \\ 1 \end{bmatrix}$$

Nilai bobot matriks yang sesuai, yaitu :

$$W^{(f)} = \begin{bmatrix} 0.11 & 0.32 \\ 0.42 & 0.19 \end{bmatrix}, W^{(i)} = \begin{bmatrix} 0.60 & 0.17 \\ 0.16 & 0.17 \end{bmatrix}, W^{(g)} = \begin{bmatrix} 0.46 & 0.74 \\ 0.75 & 0.65 \end{bmatrix},$$
$$W^{(o)} = \begin{bmatrix} 0.98 & 0.08 \\ 0.15 & 0.54 \end{bmatrix}$$

Nilai bobot matriks yang tersembunyi, yaitu :

$$U^{(f)} = \begin{bmatrix} 0.87 & 0.50 \\ 0.23 & 0.67 \end{bmatrix}, U^{(i)} = \begin{bmatrix} 0.30 & 0.89 \\ 0.64 & 0.65 \end{bmatrix}, U^{(g)} = \begin{bmatrix} 0.60 & 0.12 \\ 1.00 & 0.01 \end{bmatrix},$$
$$U^{(o)} = \begin{bmatrix} 0.41 & 0.62 \\ 0.62 & 0.14 \end{bmatrix}$$

$$b^{(f)} = [0.30 \quad 0.1], b^{(i)} = [0.67 \quad 0.13], b^{(g)} = [0.47 \quad 0.07],$$
$$b^{(o)} = [0.75 \quad 0.09]$$

Forward Propagation

1. Pertama-tama, mari kita hitung t_0 menggunakan Persamaan (1) dan Persamaan (4) yang didefinisikan sebagai berikut :

   $g_t = \widetilde{C}_t = tanh\,(W^{(g)}x_t + U^{(g)}h_{t-1} + b^{(g)})$……..……...(4)

   Perhitungan *step* pertama, yaitu :

   g_0

   $$= \tanh\left(\begin{bmatrix}0.46 & 0.75\\ 0.74 & 0.65\end{bmatrix}\cdot\begin{bmatrix}0.25\\ 0.3\end{bmatrix}+\begin{bmatrix}0.6 & 1.00\\ 0.12 & 0.01\end{bmatrix}\cdot\begin{bmatrix}0\\ 0\end{bmatrix}+\begin{bmatrix}0.47 & 0.07\end{bmatrix}\right)=\begin{bmatrix}0.66959\\ 0.42190\end{bmatrix};$$

   Atau

   $$g_0 = \tanh\left(\begin{bmatrix}0.46 & 0.75 & 0.6 & 1.00 & 0.47\\ 0.74 & 0.65 & 0.12 & 0.01 & 0.07\end{bmatrix}\cdot\begin{bmatrix}0.25\\ 0.3\\ 0\\ 0\\ 1\end{bmatrix}\right)=\tanh\left(\begin{bmatrix}0.81\\ 0.45\end{bmatrix}\right)=\begin{bmatrix}0.66959\\ 0.42190\end{bmatrix};$$

   Dan serupa dalam menghitung,

   $$f_0 = \begin{bmatrix}0.61147\\ 0.55897\end{bmatrix}, i_0 = \begin{bmatrix}0.70432\\ 0.55564\end{bmatrix}, o_0 = \begin{bmatrix}0.73885\\ 0.56758\end{bmatrix}.$$

2. *Memory state* t_0 adalah $C_0 = \begin{bmatrix}0.47161\\ 0.23442\end{bmatrix}$.

3. *Cell output* pada t_0 adalah $h_0 = \begin{bmatrix}0.32472\\ 0.13067\end{bmatrix}$.

4. Selanjutnya, setelah mengulangi langkah-langkah tersebut untuk *timestep* t_1,

   $$f_1 = \begin{bmatrix}0.74248\\ 0.69481\end{bmatrix};\ i_1 = \begin{bmatrix}0.82911\\ 0.69219\end{bmatrix};\ o_1 = \begin{bmatrix}0.88740\\ 0.69463\end{bmatrix};\ g_1 = \begin{bmatrix}0.95231\\ 0.87876\end{bmatrix}.$$

Maka, hasil $C_1 = \begin{bmatrix} 1.13973 \\ 0.77115 \end{bmatrix}$, $h_1 = \begin{bmatrix} 0.72263 \\ 0.44984 \end{bmatrix}$.

2.7.Tokenisasi

Tokenisasi merupakan salah satu langkah penting dalam melakukan *preprocessing* teks dengan membagi frasa, kalimat, paragraf, satu atau beberapa dokumen teks menjadi unit yang lebih kecil yang sebut token (Vijayarani & Janani, 2016). Token bisa berupa apa saja seperti kata, subword, atau bahkan karakter tergantung dengan algoritma dalam melakukan proses tokeninasi. Proses tokenisasi memiliki tujuan, yaitu memecah data yang tidak terstruktur menjadi potongan-potongan informasi numerik yang cocok untuk pembelajaran mesin. Tokenisasi yang digunakan untuk data protein *sequence* dalam penelitian ini adalah algoritma tokenisasi berbasis karakter, yaitu proses tokenisasi dengan membagi sebuah kalimat menjadi karakter individu. Contoh tokenisasi berbasis karakter untuk protein *sequence* dapat dilihat pada Tabel 3.

Tabel 3. Implementasi Tokenisasi Berbasis Karakter

Sebelum Tokenisasi	Tokenisasi
‘A’, ‘C’, ‘D’, ‘E’, ‘F'', ‘G’, ‘H’, ‘I’, ‘K’, ‘L’, ‘M’, ‘N’, ‘P’, ‘Q’, ‘R’, ‘S’, ‘T’, ‘V’, ‘W’, ‘Y’	‘A : 1’, ‘C : 2’, ‘D : 3’, ‘E : 4’, ‘F : 5’, ‘G : 6’, ‘H : 7’, ‘I : 8’, ‘K : 9’, ‘L : 10’, ‘M : 11’, ‘N : 12’, ‘P : 13’, ‘Q : 14’, ‘R : 15’, ‘S : 16’, ‘T : 17’, ‘V : 18’, ‘W : 19’, ‘Y : 20’

Dari tokenisasi di atas diberikan data sequence, hasilnya dapat dilihat pada Tabel 4.

Tabel 4. Hasil Tokenisasi Protein *Sequence*

Protein Sequence	Hasil Tokenisasi
PHPESRIRLSTRRDAHGMPI	[13,7,13,3,16,15,8,15,10,16,17,15,15,3,1,7,6,11,13,8]
MMNSKIAEVVVLNCRTCTRA	[11,11,12,16,9,8,1,4,18,18,18,10,12,2,15,17,2,17,15,1]

2.8.*Padding*

Padding merupakan sebuah proses mengubah setiap *sequence* agar memiliki panjang yang sama. Pada *padding*, setiap *sequence* dibuat sama panjang dengan menambahkan nilai 0 secara sufiks atau prefiks hingga mencapai panjang maksimum *sequence*. Selain itu *padding* juga dapat memotong *sequence* hingga panjangnya sesuai dengan panjang maksimum *sequence*. Hasil setelah *padding* adalah setiap *sequence* memiliki panjang yang sama. *Padding* dapat melakukan ini dengan menambahkan 0 secara *default* pada awal *sequence* yang lebih pendek. Nilai 0 yang diisi pada awal *sequence* disebut *pre-padding* dan nilai 0 yang diisi pada akhir *sequence* disebut *post-padding.* Contoh penggunaan *padding* dapat dilihat pada Tabel 5.

Table 5. Contoh Penggunaan *Padding*

Sequence	***Pre-padding***	***Post-padding***
[2,3,4,5]	[2,3,4,5]	[2,3,4,5]
[3,4]	[0,0,3,4]	[3,4,0,0]

2.9.*Embedding layer*

Proses mengubah kata *input* menjadi sebuah vektor yang digunakan dalam pemrosesan bahasa alami merupakan fungsi dari *embedding layer* (Khrulkov, et al., 2019). *Embedding layer* didefinisikan sebagai *hidden*

layer pertama dari sebuah jaringan. *Input* diskrit yang diubah menjadi banyak titik-titik vektor di dalam sebuah ruang vektor disebut dengan *embedding vector*. *Embedding vector* merupakan vektor yang mewakili kata-kata dalam ruang L-dimensi, dimana L adalah panjang vektor.

2.10. *Random Undersampling*

Random undersampling adalah metode efisien yang sering digunakan dalam mengklasifikasi *imbalance* data atau data yang tidak seimbang (Ganganwar, 2012). Teknik ini menghapus contoh dari kelas mayoritas, yaitu kelas yang lebih banyak sehingga jumlahnya sama dengan kelas minoritas, yaitu kelas yang jumlahnya lebih sedikit. *Random undersampling* mengacu pada proses pengurangan jumlah sampel. Sampel dari kelas mayoritas dipilih secara acak dengan atau tanpa penggantian. Setelah pengambilan sampel secara acak, jumlah kelas mayoritas dalam kumpulan data berkurang, sehingga mengurangi waktu pelatihan model secara signifikan.

2.11. *Confusion Matrix*

Confusion matrix adalah sebuah tabel yang sering digunakan untuk mengukur kinerja dari model klasifikasi**.** *Confusion matrix* mencatat jumlah kejadian dari klasifikasi yang diprediksi (Heydarian, Doyle & Samavi, 2022)**.** Perhitungan *Confusion Matrix* dapat dilihat pada Tabel 6.

Tabel 6. *Confusion Matrix* pada Klasifikasi Dua Kelas

	Actual True	***Actual False***
Predicted True	*True Positive* (TP)	*False Positive* (FP)
Predicted False	*False Negative* (FN)	*True Negative* (TN)

Berdasarkan Tabel 6, yang diidentifikasi dengan benar oleh algoritma sebagai positif disebut *true positive* (TP), dan diidentifikasi salah diklasifikasikan sebagai negatif diberi label *false negative* (FN). Di sisi lain, elemen negatif yang benar berlabel negatif disebut *true negative* (TN), sedangkan yang salah prediksi sebagai positif disebut *false positive* (FP) (Chicco, Tötsch & Jurman, 2021). Beberapa matriks pengukuran yang digunakan untuk mengevaluasi kinerja prediksi model pada penelitian ini, yaitu :

a. *Sensitivity* (SE)

   Sensitivity adalah nilai kelengkapan atau keakuratan kelas positif yang diberi label dengan benar (Bekkar, Djemaa & Alitouche, 2013). Rumus *sensitivity* dapat dilihat pada Persamaan (5).

   $$SN = \frac{TP}{TP+FN} \quad \text{.....(5)}$$

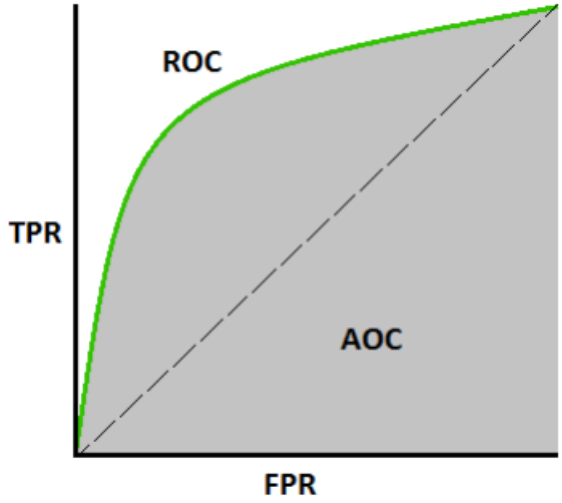
b. *Specificity* (SP)

   Specificity adalah probabilitas bersyarat dari kelas negatif yang diberi label dengan benar (Bekkar, Djemaa & Alitouche, 2013). Rumus *Specificity* dapat dilihat pada Persamaaan (6).

   $$SP = \frac{TN}{TN+FP} \quad \text{.....(6)}$$

2.12. ROC-AUC

Kurva AUC - ROC adalah pengukuran kinerja untuk masalah klasifikasi pada berbagai pengaturan ambang batas. Kurva *Receiver Operating Characteristic* (ROC) adalah kurva probabilitas dan kurva *Area Under Curva* (AUC) adalah kurva yang mewakili ukuran seluruh area dua dimensi di bawah kurva ROC. Semakin tinggi AUC, maka semakin baik model memprediksi 0 sebagai 0 dan 1 sebagai 1. Kurva ROC-AUC dapat dilihat pada Gambar 4.



Gambar 4. Kurva ROC-AUC (Narkhede, 2019).

Model yang sangat baik memiliki AUC mendekati 1 yang berarti memiliki ukuran keterpisahan yang baik. Model yang buruk memiliki AUC mendekati 0 yang berarti memiliki ukuran keterpisahan yang paling buruk dan ketika AUC adalah 0,5, berarti model tidak memiliki kapasitas pemisahan kelas apa pun. Kurva *Receiver Operating Characteristic* (ROC) diplot dengan TPR terhadap FPR di mana TPR berada pada sumbu y dan FPR berada pada sumbu x. (Narkhede, 2019). TPR dan FPR dirumuskan seperti berikut :

1. *True Positive Rate* (TPR) dirumuskan seperti pada Persamaan (7).

$$\text{TPR} = \frac{TP}{TP+FN} \quad \text{.....(7)}$$

TPR diperoleh dari nilai *true positive* yang dibagi dengan jumlah *true positive* dan *false negative.*

2. *False Positive Rate* (FPR) dirumuskan seperti Persamaan (8).

$$\text{FPR} = \frac{FP}{FP+TN} \quad \text{.....(8)}$$

FPR diperoleh dari nilai *false negative* yang dibagi dengan jumlah *true positive* dan *false negative.*

2.13. PR-AUC

Kurva *precision-recall* dibangun dengan menghitung dan memplot *precision* terhadap *recall* untuk satu *classifier*.Kurva *precision-recall* juga telah diakui berguna untuk penilaian kinerja klasifikasi untuk respons biner yang tidak seimbang dalam bioinformatika (Saito & Rehmsmeier, 2015). *Precision* merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan hasil yang diprediksi positif. Perhitungan *precision* ditunjukkan oleh Persamaan (9).

$$Precision = \frac{TP}{TP+FP} \quad \text{.....(9)}$$

Recall merupakan rasio prediksi benar positif dibandingkan dengan keseluruhan data yang benar positif. Perhitungan *recall* ditunjukkan oleh Persamaan (10).

$$Recall = \frac{TP}{TP+FN} \quad \text{.....(10)}$$

III. METODOLOGI PENELITIAN

3.1. Tempat dan Waktu Penelitian

Berikut adalah penjelasan mengenai tempat dan waktu pada saat penelitian:

3.1.1. Tempat Penelitian

Penelitian dilaksanakan di Lab Rekayasa Perangkat Lunak (RPL) yang berada di Gedung MIPA Terpadu Jurusan Ilmu Komputer Fakultas Matematika dan Ilmu Pengetahuan Alam Universitas Lampung.

3.1.2. Waktu dan Jadwal Penelitian

Penelitian dilakukan pada bulan Desember 2022 di semester tujuh ganjil hingga penyelesaian pada bulan Agustus 2023. Alur pengerjaan penelitian dibagi menjadi 3 tahap, yaitu:

1. Tahap Perancangan Penelitian

   Tahap pertama yang dilakukan dalam perancangan penelitian, yaitu pengumpulan data. Pada tahap ini juga dilakukan pemahaman studi literatur, menentukan metode yang digunakan untuk penelitian dan melakukan penyususan draft Bab 1-3. Data didapatkan melalui jurnal atau paper yang menjadi acuan setelah melakukan studi literatur sebelumnya.

2. Tahap penelitian lanjutan

   Tahap yang selanjutnya pelaksanaan penelitian, yaitu melakukan *preprocessing* data terhadap *sequence* yang akan diolah. Selanjutnya, melakukan melakukan pemodelan klasifikasi dan prediksi menggunakan metode *Long Short-Term Memory* (LSTM).

3. Tahap evaluasi

   Tahap yang terakhir, yaitu tahap evaluasi, pada tahap ini melakukan penulisan *draft* hasil untuk Bab 4-5 yang digunakan dalam penyampaian hasil penelitian melalui seminar hasil.

3.2.Data dan Alat

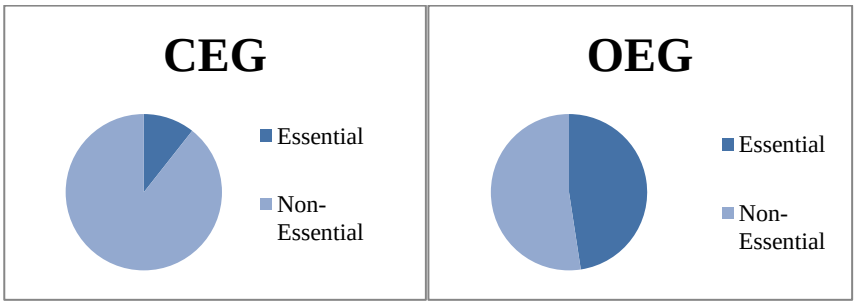
3.2.1. Data

Data yang diambil merupakan data yang diambil dari OGEE (*Online GEne* Esensial*ity database*) dan DEG (*Database of* Esensial *Genes*) oleh penelitian terdahulu (Beder, et al., 2021).. Dataset terdiri dari dataset CEG (*Celuller* Esensial *Gene*) dan OEG (*Organismal* Esensial *Gene*). Dataset CEG merupakan dataset gen esensial yang terlibat dalam proses biogenesis, makromolekul seluler dan siklus sel atau poliferasi sedangkan OEG adalah dataset gen esensial yang terlibat dalam proses pengayaan dalam regulasi, perkembangan atau morfogenesis, proses yang terkait saraf serta persinyalan. Jumlah data dapat dilihat pada Tabel 7.

Table 7. Data *Sequence* Protein *Drosophila Melanogaster*

Jenis Data	Esensial	*Non*-esensial	Jumlah
CEG	1.227	10.320	11.547
OEG	246	271	517

Berdasarkan Tabel 7, persentase data gen esensial dan non-esensial pada dataset CEG dan OEG dapat dilihat pada Gambar 5.



Gambar 5. Persentase Perbandingan Esensial dan *Non*-Esensial.

Esensial gen diidentifikasikan menjadi 2 kategori, yaitu CEG dan OEG. Untuk CEG terdiri dari 1.227 gen esensial dan 10.320 gen *non*-esensial, sedangkan OEG terdiri dari 246 gen esensial dan 271 gen *non*-esensial. Jadi jumlah keseluruhan data yang dipakai adalah 11.547 untuk data CEG dan 517 untuk data OEG. Sequence terpendek pada CEG adalah berjumlah 60 dan terpanjang adalah 20.710. Panjang sequnce terpendek pada OEG adalah 73 dan terpanjang adalah 16.236. Berikut bentuk dari dataset yang didapatkan dapat dilihat pada Gambar 6.

	protein	gene
0	MAVRYELAIGLNKGHKTSKIRNVKYTGDKKVKGLRGSRLKNIQTRH...	Essential
1	MEPIGDLQVPSFKVVSGGTTFTYASPKSGAASLDFLAHTLRKREAN...	Non-essential
2	MMNSKIAEVVVLNCRTCTRACKLHKPLQEEIDLGSEGSTTLASMLN...	Essential
3	MSQESNGGPAAGGGAAAAPPPPPQYIITTPSEVDPDEVRSMADLEL...	Essential

Gambar 6. Dataset Protein *Sequence*.

3.2.2. Alat

3.2.2.1.Perangkat Keras (*Hardware*)

Perangkat keras (*hardware*) yang digunakan dalam penelitian ini, yaitu :

a. *Processor* : *Intel*(R) *Core*(TM) i5-8250U CPU @ 1.60GHz, 1800 Mhz, 4 *Core*(s), 8 *Logical Processor*(s)
b. *Random Access Memory* (RAM) : 4.00 GB,
c. *Storage* : HDD 1 TB
d. *Video Graphics Array* (VGA) : AMD Radeon™ 530

3.2.2.2.Perangkat Lunak (*Software*)

Perangkat lunak (*software*) yang digunakan dalam penelitian ini, yaitu :

a. *Operating System* : Windows 10 Home Single Language 64-bit
b. Microsoft Excel 2010

c. *Tools*
   a) Jupyter Notebook adalah *tool* yang dipakai untuk mengolah data di python.
   b) Google Colab

      Google Colab atau Google Colaboratory, adalah sebuah *executable document* yang dapat digunakan untuk menyimpan, menulis, serta membagikan program yang telah ditulis melalui Google Drive. *Software* ini pada dasarnya serupa dengan Jupyter Notebook gratis berbentuk *cloud* yang dijalankan menggunakan *browser,* seperti Mozilla Firefox dan Google Chrome.
   c) Python 3.9.12

Python merupakan bahasa pemrograman tinggi yang bisa melakukan eksekusi sejumlah instruksi multi guna secara langsung (interpretatif) dengan metode *Object Oriented Programming* dan juga menggunakan semantik dinamis untuk memberikan tingkat keterbacaan *syntax*.

d. Packages

a) Library Pandas 1.4.2

Pandas adalah sebuah *library* berlisensi BSD dan *open source* yang menyediakan struktur data dan analisis data yang mudah digunakan dan berkinerja tinggi untuk bahasa pemrograman python. Pandas melakukan tugas penting seperti menyelaraskan data untuk perbandingan dan penggabungan set data.

b) Library Sklearn 1.0.2

Scikit-learn adalah salah satu *package* yang memudahkan untuk melakukan *processing* data. *Package* ini menyediakan pilihan alat yang efisien klasifikasi, regresi, *clustering*, dan *dimension reduction* melalui antarmuka konsistensi dengan python

c) Numpy 1.21.5

Numpy merupakan salah satu *library* python yang banyak digunakan dalam proses analisis data. Numpy adalah sebuah *package* yang bekerja pada bahasa pemograman python. numpy biasa digunakan untuk mengolah data numerik/saintifik. Data yang diolah dapat berupa multidimensional *array*.

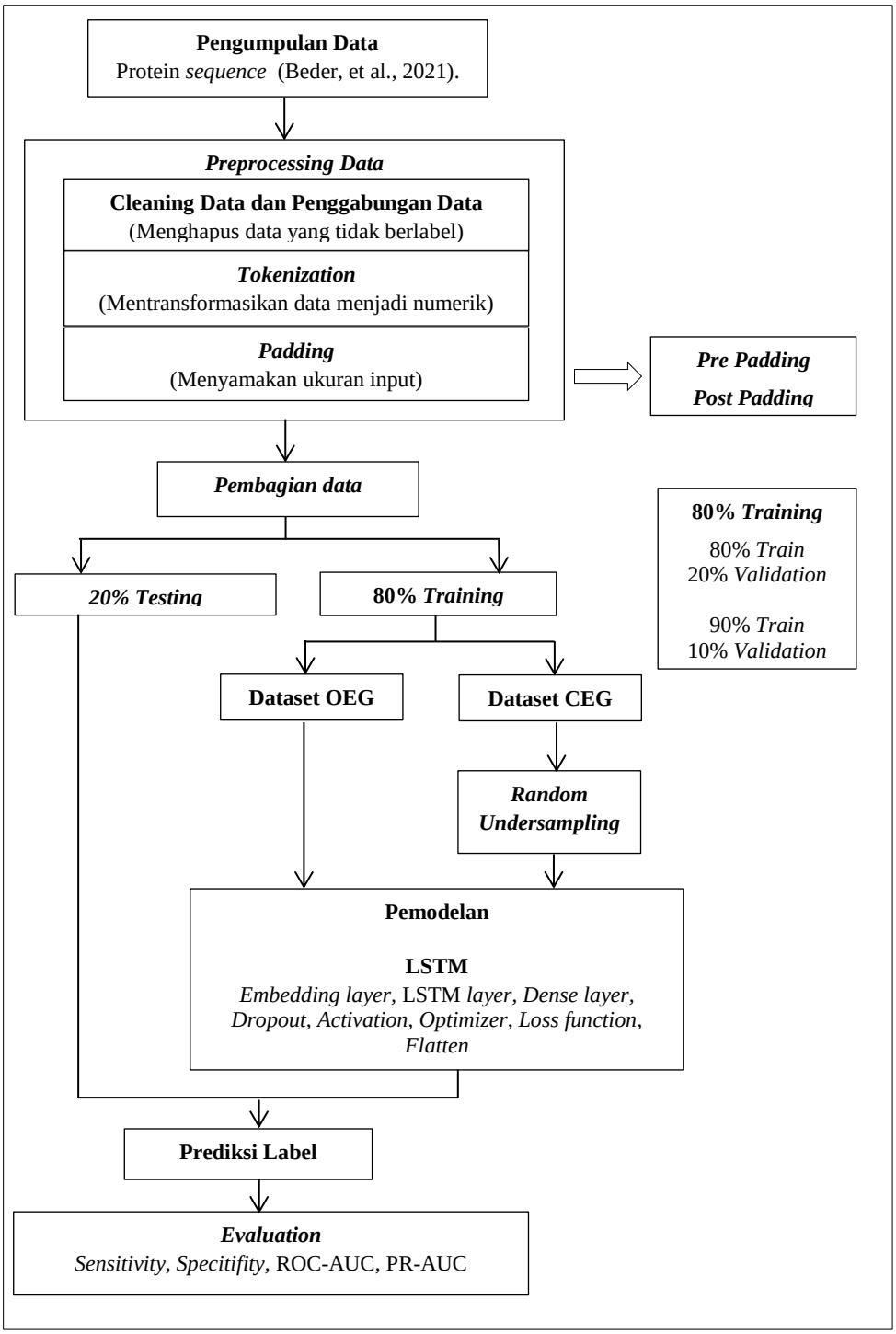
d) *Tensorflow* 2.9.1

Tensorflow adalah *library* yang buat oleh google untuk melakukan pemodelan *deep learning* dengan

mudah. *Tensorflow* merupakan *library* yang berjalan pada bahasa pemrograman yang lebih rendah (mendekati bahasa mesin) sehingga memiliki performa komputasi yang baik.

3.3.Metodologi

Alur kerja penelitian klasifikasi gen esensial pada *Drosophila melanogaster* dapat dilihat pada Gambar 7.



Gambar 7. Alur Kerja Penelitian Klasifikasi Gen Esensial pada *Drosophila Melanogaster* Menggunakan Metode LSTM.

3.3.1. Pengumpulan Data

Data *sequence* protein dari *Drosophila melanogaster* diperoleh dari penelitian Beder, et al. (2021). Data tersebut terdiri dari 2 kategori, yaitu CEG dan OEG. Untuk CEG terdiri dari 1.227 gen esensial dan 10.320 gen *non*-esensial, sedangkan OEG terdiri dari 246 gen esensial dan 271 gen *non*-esensial. Jadi jumlah keseluruhan data yang dipakai adalah 11.547 untuk data CEG dan 517 data OEG.

3.3.2. *Preprocessing* Data

Pada penelitian ini menggunakan data yang berbentuk *text* karena berupa *sequence* protein. Setelah data didapatkan tahap selanjutnya, yaitu proses *cleaning* data agar dapat digunakan dengan menyisakan data-data yang memiliki label, kemudian melakukan penggabungan data dan tokenisasi dengan tipe *Char-level tokenization*. Setelah melakukan tokenisasi tahap selanjutnya, yaitu melakukan *padding* untuk menyamakan ukuran input. Pada data CEG dilakukan *teknik random undersampling* yang dilakukan setelah proses *padding.*

3.3.3. Pembagian Data

Langkah selanjutnya yang dilakukan, yaitu membagi data menjadi data *training*, *validation* dan data *testing*. Data *training* adalah data yang digunakan untuk melatih atau membangun model. Data *validation* adalah data yang digunakan untuk mengoptimasi saat melatih model atau data yang digunakan untuk menguji kinerja model pada saat *training*. Data *testing* adalah data yang digunakan untuk menguji model setelah proses *training* selesai. Data dibagi menjadi sebesar data *training* 80% dan data *testing* 20% pada masing masing dataset. Data *training* kemudian dibagi lagi menjadi data *train* dan *validation* dengan 2 skema pembagian, yaitu data *train* 80% data *validation* 20% dan data *train* 90% data *validation* 10%.

3.3.4. Model dan Klasifikasi

Pada tahap ini dilakukan pemodelan dan pengklasifikasian gen essensial menggunakan metode *Long Short-Term Memory* (LSTM). Pemodelan menggunakan beberapa layer seperti *embedding layer, LSTM layer, dense layer* dan *dropout. Layer embedding* merupakan lapisan pertama setelah melakukan proses *padding*. Dalam *embedding layer* terdapat 3 parameter, yaitu *input* dimensi, *output* dimensi dan *input* length. *Input* dimensi adalah ukuran kosa kata dimana pada penelitian ini memiliki 20 karakter asam amino dan karakter *padding* (0) sehingga input dimensinya adalah 21. *Output* dimensi adalah panjang vektor dalam setiap karakter. *Input length* adalah panjang maksimum urutan, panjang maksimum urutan ini mengikuti jumlah panjang maksimum pada *padding*. Pada penelitian ini merancang 2 arsitektur. Pada lapisan LSTM untuk arsitektur I neuron yang digunakan adalah 64 dan untuk arsitektur II neuron yang digunakan adalah 32.

3.3.5. *Testing* dan Evaluasi

Tahap selanjutnya, yaitu menghitung nilai pengujian atau evaluasi yang menghasilkan nilai *confusion matrix* dengan nilai *sensitivity, specificity*, ROC-AUC, PR-AUC. Dalam memilih model yang baik penelitian ini memiliki indikasi, yaitu :

a. Pertama, yaitu melihat hasil PR-AUC. Hal ini dikarenakan PR-AUC memplot nilai *precision* dan *recall*, dimana *precision* mengukur keakuratan prediksi positif dan *recall* mengukur kemampuan untuk mengidentifikasi kejadian positif dengan benar. Dengan menggunakan PR-AUC, model akan berfokus pada penilaian kelas positif dibandingkan kelas negatif.
b. Kedua, yaitu mempertimbangkan nilai ROC-AUC dalam mengevaluasi model klasifikasi, dimana ROC memplot *True Positive Rate* (TPR), yaitu probabilitas bahwa sampel positif

diprediksi dengan benar di kelas positif dan FPR, yaitu probabilitas bahwa sampel negatif salah diprediksi di kelas positif. Metrik ini baik digunakan untuk distribusi kelas yang seimbang. FPR dianggap baik jika nilai yang dihasilkan kecil, karena menunjukkan lebih sedikit kesalahan positif. Namun pada data yang tidak seimbang, FPR cenderung tetap pada nilai yang kecil karena banyaknya angka negatif (membuat penyebut menjadi besar).

c. Ketiga mempertimbangkan nilai *sensitivity* dan *specificity*. ROC-AUC dapat menyebabkan kinerja yang kurang tepat pada dataset yang tidak seimbang, maka penting untuk melihat persentase kelas positif dan negatif yang benar diprediksi untuk memastikan bahwa model yang dipilih dapat memprediksi kelas dengan baik selain dengan melihat hasil dari PR-AUC.

V. PENUTUP

5.1.Kesimpulan

Pada penelitian ini dapat diambil kesimpulan sebagai berikut.

1. Pada penelitian ini yang mengklasifikasikan gen esensial pada protein *sequence* dari organisme *Drosophila Melanogaster* (Lalat Buah) menggunakan metode *Long Short-Term Memory* mendapatkan hasil klasifikasi terbaik untuk dataset OEG dengan pembagian data 80% *training* dan 20% validasi pada arsitektur II menggunakan *pre padding* memiliki hasil yang paling tinggi diantara hasil yang lainnya, dengan nilai *sensitivity* 81%, *specificity* 76%, ROC-AUC 79% , PR-AUC 82%. Pada dataset CEG hasil terbaik dengan pembagian data 80% *training* dan data 20% validasi pada arsitektur I menggunakan *post-padding* mendapatkan nilai akurasi tertinggi, dengan *sensitivity* 73%, *specitifity* 50% ROC-AUC 61%, dan PR-AUC 45%.
2. Hasil pengujian pada penelitian ini mendapatkan hasil lebih rendah dari penelitian sebelumnya (Beder, et al., 2021). Hal ini berarti metode *Long Short-Term Memory* pada penelitian ini, belum cukup baik dalam mengklasifikasikan protein pada *Drosophila melanogaster* dengan parameter yang digunakan.

5.2.Saran

Adapun saran yang diberikan pada penelitian ini adalah sebagai berikut.

1. Penelitian ini dapat menggunakan metode klasifikasi lainnya untuk mendapatkan hasil yang lebih baik.
2. Penelitian ini dapat dilanjutkan dengan menggunakan parameter lain untuk membangun kinerja model agar mendapatkan hasil yang lebih baik.

DAFTAR PUSTAKA

Houdt, G. V., Mosquera, C., & Nápoles, G. (2020). A Review on the Long Short-Term Memory Model. *Artificial Intelligence Review*.

Khanh Le, N. Q., Do, D. T., Hung, T. K., Lam, L. T., Huynh, T.-T., & Nguyen, N. K. (2020). A Computational Framework Based on Ensemble Deep Neural Networks for Essential Genes Identification. *Molecular Science*.

Khrulkov, V., Hrinchuk, O., Mirvakhabova, L., & Oseledets, I. (2019). Tensorized Embedding Layers for Efficient Model Compression. *ICML*.

Adams, M. D., Holt, R. A., Li, P. W., & Richards, S. (2000). The Genome Sequence of Drosophila melanogaster. *Science*, 2185-2195.

Alasadi, S., & Bhaya, W. (2017). Review of Data Preprocessing Techniques in Data Mining. *Journal of Engineering and Applied Sciences*, 4102 - 4107.

Aromolaran, O., Beder, T., Oswald, M., Oyelade, J., Adebiyi, E., & Koenig, R. (2020). Essential gene prediction in Drosophila melanogaster using machine learning approaches based on sequence and functional features. *Computational and Structural Biotechnology Journal*, 612-621.

Azhar, M. (2016). *BIOMOLEKUL SEL Karbohidrat, Protein, dan Enzim*. Padang: UNP Press.

Beder, T., Aromolaran, O., Donitz, J., Tapanelli, S., Adedeji, E., Adebiyi, E. (2021). Identifying Essential Genes Across Eukaryotes by Machine Learning. *NAR Genomics and Bioinformatics*.

Bekkar, M., Djemaa, D. K., & Alitouche, D. A. (2013). Evaluation Measures for Models Assessment over Imbalanced Data Sets. *Journal of Information Engineering and Applications*.

Branden, C., & Tooze, J. (2012). *Introduction to Protein Structure.* Garland Science.

Campos, T. L., Korhonen, P. K., Hofmann, A., Gasser, R. B., & Young, N. D. (2020). Combined use of feature engineering and machine-learning to predict essential genes in Drosophila melanogaster. *NAR Genomics and Bioinformatics*.

Campos, T. L., Kornohen, P. K., Gasser, R. B., & Young, N. D. (2019). An Evaluation of Machine Learning Approaches for the Prediction of Essential Genes in Eukaryotes Using Protein Sequence-Derived Features. *Computational and Structural Biotechnology*, 785 - 796.

Cerniker, S. E., & Rubin, G. M. (2003). The Drosophila Melanogaster Genome. *Annu Rev Genomics Hum Genet*, 89-117.

Chicco, D., Tötsch, N., & Jurman, G. (2021). The Matthews Correlation Coefficient (MCC) is More Reliable than Balanced Accuracy, Bookmaker Informedness, and Markedness in Two-Class Confusion Matrix Evaluation. *BioData Mining* .

Fairuz, A. Z., Afifah, Fahrizal, M., Annisa, N., & Sari, T. R. (2022). Metabolisme Protein Dalam Tubuh Manusia. *Jurnal Ilmu Alam Indonesia*.

Ganganwar, V. (2012). An Overview of Classification Algorithms for Imbalanced Datasets. *IJETAE*, 43-47.

Heydarian, M. R., Doyle, T. E., & Samavi, R. (2022). MLCM: Multi-Label Confusion Matrix. *IEEE Access*.

Liu, X., Wang, B.-J., Xu, L., Tang, H.-L., & Xu, G.-Q. (2017). Selection of Key Sequence-Based Features for Prediction of Essential Genes in 31 Diverse Bacterial Species. *PLOS ONE*.

McLaughlin , J. M., & Bratu, D. P. (2015). *Drosophila melanogaster Oogenesis: An Overview*. New York: Springer Science.

Miklos, G. G., & Rubin, G. M. (1996). The Role of the Genome Project in Determining Gene Function: Insights from Model Organisms. *Department of Molecular and Cell Biology*, 521-529.

Nainu, F. (2018). Review : Penggunaan Drosophila melanogaster Sebagai Organisme Model Dalam Penemuan Obat. *Jurnal Farmasi Galenika (Galenika Journal of Pharmacy)*, 50-67.

Narkhede, S. (2019). *Understanding AUC - ROC Curve.* Towards Data Science.

O'Grady, P. M., & Markow, T. A. (2009). Phylogenetic taxonomy in Drosophila. *Landes Bioscience*, 10-14.

Peng, C., Lin, Y., Luo, H., & Gao, F. (2017). A Comprehensive Overview of Online Resources to Identify and Predict Bacterial Essential Genes. *Frontiers in Microbiology*.

Perveen, F. K. (2017). *Drosophila melanogaster: Model for Recent Advances in Genetics and Therapeutics.* InTech.

Saidi, R., Aridhi, S., Nguifo, E. M., & Maddouri, M. (2012). Feature Extraction in Protein Sequences Classification: a new stability measure. *Computational Biology and Biomedicine*, 683–689.

Saito, T., & Rehmsmeier, M. (2015). The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets. *Computational Biology Unit*.

Salma, N., & Rustam, Z. (2019). Naïve Bayes Classifier Models for Predicting the Colon Cancer. *Materials Science and Engineering*, 546.

Sharma, A. K., Eils, R., & Konig, R. (2016). Copy Number Alterations in Enzyme-Coding and Cancer-Causing Genes Reprogram Tumor Metabolism. *Integrated Systems and Technologies*, 4058–4067.

Smagulova, K., & James, A. P. (2020). Overview of Long Short-Term Memory Neural Networks. *Modeling and Optimization in Science and Technologies* .

Steinmetz, L. M., Scharfe, C., Deutschbauer, A. M., Mokranjac, D., Herman, Z. S., Jones, T. (2002). Systematic Screen for Human Disease Genes in Yeast. 400 – 404.

Suprayitno, E., & Sulistiyati, T. D. (2017). *Metabolisme Protein.* Malang: UB Press.

Vijayarani, S., & Janani, R. (2016). Text Mining: Open Sourch Tokenization Tools – an Analysis. *Advanced Computational Intelligence: An International Journal (ACII)*, 37-47.

Xingyi Li, Wenkai Li, Min Zeng, Ruiqing , Z., & Min Li. (2019). Network-based methods for predicting essential genes or proteins: a survey. *Briefings in Bioinformatics*, 566–583.

Xu, P., Ge, X., Chen, L., Wang, X., Duo, Y., Xu, J. (2011). Genome-Wide Essential Gene Identification in Streptococcus Sanguinis. *SCIENTIFIC REPORTS*, 1:125.

Zhang, R., Ou, H.-Y., & Zhang, C.-T. (2004). DEG: a Database of Essential Genes. *Nucleic Acids Research*, 271-272.

Zhang, Z., & Ren, Q. (2015). Why are essential genes essential? - The essentiality of Saccharomyces genes. *Microbial Cell*, 280-287.