

Embedding Struktur Sekunder IncRNA Berbasis Variational Autoencoder untuk Klasifikasi Kanker Hati dengan GraphSAGE

Theodora Tantri Trisnawati dan Mohammad Isa Irawan



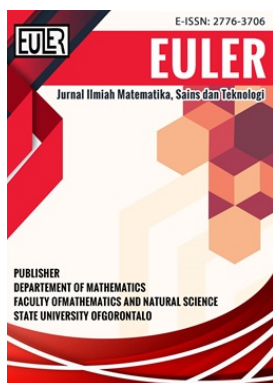
Volume 14, Issue 2, Pages 593–609, August 2026

Diterima 23 Juni 2026, Direvisi 21 Agustus 2026, Disetujui 23 Agustus 2026, Diterbitkan 27 Agustus 2026

To Cite this Article : T. T. Trisnawati dan M. I. Irawan, “Embedding Struktur Sekunder IncRNA Berbasis Variational Autoencoder untuk Klasifikasi Kanker Hati dengan GraphSAGE”, Euler J. Ilm. Mat. Sains dan Teknol., vol. 14, no. 2, pp. 593–609, 2026, doi: <https://doi.org/10.37905/euler.v14i2.40387>

© 2026 by author(s)

JOURNAL INFO • EULER : JURNAL ILMIAH MATEMATIKA, SAINS DAN TEKNOLOGI

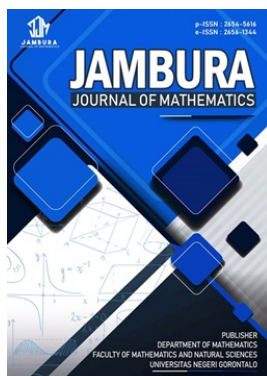


	Homepage	:	http://ejurnal.ung.ac.id/index.php/euler/index
	Journal Abbreviation	:	Euler J. Ilm. Mat. Sains dan Teknol.
	Frequency	:	Three times a year
	Publication Language	:	English (preferable), Indonesia
	DOI	:	https://doi.org/10.37905/euler
	Online ISSN	:	2776-3706
	License	:	Creative Commons Attribution-NonCommercial 4.0 International License
	Publisher	:	Department of Mathematics, Universitas Negeri Gorontalo
	Country	:	Indonesia
	OAI Address	:	http://ejurnal.ung.ac.id/index.php/euler/oai
	Google Scholar ID	:	QF_r-gAAAAJ
	Email	:	euler@ung.ac.id

JAMBURA JOURNAL • FIND OUR OTHER JOURNALS



Jambura Journal of Biomathematics



Jambura Journal of Mathematics



Jambura Journal of Mathematics Education



Jambura Journal of Probability and Statistics

Embedding Struktur Sekunder lncRNA Berbasis Variational Autoencoder untuk Klasifikasi Kanker Hati dengan GraphSAGE

Theodora Tantri Trisnawati¹, Mohammad Isa Irawan^{1,*}

¹Departemen Matematika, Institut Teknologi Sepuluh Nopember, Surabaya, 60111, Indonesia

ARTICLE HISTORY

Diterima 23 Juni 2026
Direvisi 21 Agustus 2026
Disetujui 23 Agustus 2026
Diterbitkan 27 Agustus 2026

KATA KUNCI

VAE
GraphSAGE
lncRNA
Kanker Hati

KEYWORDS

VAE
GraphSAGE
lncRNA
Liver Cancer

ABSTRAK. Identifikasi lncRNA yang berasosiasi dengan kanker hati merupakan permasalahan klasifikasi yang dapat memanfaatkan berbagai karakteristik biologis, seperti struktur sekunder, ekspresi, serta hubungan molekuler dengan RNA-binding protein (RBP). Penelitian ini mengembangkan kerangka VAE-GraphSAGE untuk mengintegrasikan informasi tersebut dalam graf heterogen. Variational Autoencoder (VAE) digunakan sebagai metode representation learning untuk mengekstraksi embedding struktur sekunder lncRNA melalui proses rekonstruksi tanpa menggunakan label kelas. Embedding tersebut kemudian dikombinasikan dengan fitur ekspresi untuk membentuk fitur node, sementara informasi kemiripan lncRNA, interaksi lncRNA-RBP, dan protein-protein interaction (PPI) digunakan dalam pembentukan graf heterogen. Selanjutnya, GraphSAGE digunakan sebagai classifier untuk memanfaatkan informasi hubungan antar-node dalam melakukan klasifikasi lncRNA. Dataset yang digunakan terdiri atas 1.000 lncRNA, yaitu 400 lncRNA positif dan 600 lncRNA negatif, serta 85 RBP sebagai node protein, dengan pembagian data training, validasi, dan testing sebesar 70:15:15. Hasil perbandingan embedding menunjukkan bahwa VAE-GraphSAGE memiliki recall tertinggi dibandingkan AE-GraphSAGE, MLP-GraphSAGE, CNN-GraphSAGE, dan ResNet-GraphSAGE, sedangkan MLP-GraphSAGE memberikan accuracy, precision, specificity, dan F1-score tertinggi, serta CNN-GraphSAGE menghasilkan ROC-AUC tertinggi. Selanjutnya, dengan menggunakan embedding VAE yang sama, VAE-GraphSAGE menghasilkan performa lebih baik dibandingkan VAE-HGT dan VAE-RGCN pada metrik klasifikasi yang dievaluasi. Hasil tersebut menunjukkan bahwa VAE efektif dalam menghasilkan representasi struktur sekunder yang mendukung deteksi lncRNA positif, sementara GraphSAGE mampu memanfaatkan hubungan biologis pada graf heterogen secara efektif untuk proses klasifikasi.

ABSTRACT. Identifying liver cancer-associated lncRNAs is a classification problem that can leverage various biological characteristics, such as secondary structure, expression levels, and molecular associations with RNA-binding proteins (RBPs). This study develops a VAE-GraphSAGE framework to integrate this information within a heterogeneous graph. A Variational Autoencoder (VAE) is employed for representation learning to extract lncRNA secondary structure embeddings through a reconstruction process that does not rely on class labels. These embeddings are combined with expression features to form node features, while data on lncRNA similarity, lncRNA-RBP interactions, and protein-protein interactions (PPI) are used to construct the heterogeneous graph. Subsequently, GraphSAGE serves as the classifier, utilizing inter-node relationship information to classify the lncRNAs. The dataset comprises 1,000 lncRNAs (400 positive and 600 negative) and 85 RBPs (as protein nodes), with a training-validation-testing split of 70:15:15. Comparative results indicate that VAE-GraphSAGE achieves the highest recall among the tested models (outperforming AE-GraphSAGE, MLP-GraphSAGE, CNN-GraphSAGE, and ResNet-GraphSAGE); meanwhile, MLP-GraphSAGE yields the highest accuracy, precision, specificity, and F1-score, and CNN-GraphSAGE produces the highest ROC-AUC. Furthermore, using the same VAE embeddings, VAE-GraphSAGE demonstrates superior performance compared to VAE-HGT and VAE-RGCN across the evaluated classification metrics. These results demonstrate that the VAE effectively generates secondary structure representations that facilitate the detection of positive lncRNAs, while GraphSAGE successfully leverages biological relationships within the heterogeneous graph for the classification task.



This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution-NonCommercial 4.0 International License. **Editorial of EULER:** Department of Mathematics, Universitas Negeri Gorontalo, Jln. Prof. Dr. Ing. B. J. Habibie, Bone Bolango 96554, Indonesia.

*Penulis Korespondensi.

1. Pendahuluan

Kanker hati merupakan salah satu penyakit paling mematikan di dunia. Berdasarkan data statistik tahun 2022, kanker hati menempati urutan ketiga sebagai penyebab kematian akibat kanker secara global dengan tingkat mortalitas mencapai 87,6% dari seluruh kasus yang tercatat. Asia merupakan wilayah dengan jumlah pasien tertinggi. Dari 607.361 kasus yang dilaporkan, sebanyak 530.928 kasus berakhir dengan kematian. Indonesia juga menunjukkan kondisi yang serupa, dengan tercatat 242.988 kematian akibat kanker hati [1]. Lebih mengkhawatirkan lagi, kanker hati diperkirakan mencakup hingga 80% dari seluruh diagnosis penyakit hati di dunia. Jumlah kasus maupun angka kematian diproyeksikan meningkat hingga 50% dalam dua dekade mendatang apabila tidak terjadi penurunan angka kejadian minimal 3% setiap tahun [2, 3].

Tingginya angka kematian tersebut berkaitan erat dengan kompleksitas kondisi klinis yang menyertai kanker hati, salah satunya adalah *acute-on-chronic liver failure*, yaitu kegagalan hati akut yang terjadi pada pasien dengan penyakit hati kronis yang telah ada sebelumnya. Berbagai upaya terapi telah dikembangkan, mulai dari terapi suportif medis, terapi hati buatan (*artificial liver therapy*), hingga transplantasi hati. Meskipun demikian, tingkat mortalitas akibat kanker hati masih tetap tinggi. Hal ini disebabkan karena mekanisme patogenesis kanker hati hingga saat ini belum sepenuhnya dipahami. Penelitian terbaru menunjukkan bahwa perkembangan kanker hati melibatkan cedera langsung pada jaringan hati maupun cedera yang dimediasi oleh sistem imun sehingga dapat menyebabkan kegagalan multiorgan [4]. Selain kompleksitas mekanisme penyakitnya, tantangan besar lainnya adalah sulitnya mendeteksi kanker hati pada stadium awal. Gejala awal seperti penurunan nafsu makan, penurunan berat badan, kelemahan, kelelahan, dan penyakit kuning bersifat tidak spesifik sehingga sering diabaikan. Akibatnya, sebagian besar kasus baru terdiagnosis ketika telah memasuki stadium lanjut, di mana pilihan terapi menjadi sangat terbatas [5].

Kondisi tersebut mendorong berbagai penelitian untuk mengembangkan metode diagnosis yang lebih akurat dan sensitif, mulai dari analisis kelainan jaringan hati, analisis ikatan molekuler, hingga pengamatan perubahan perilaku biologis pada tingkat molekuler. Oleh karena itu, penemuan *biomarker* yang sangat spesifik dan mampu mendeteksi kanker hati pada stadium dini menjadi kebutuhan yang sangat mendesak. Salah satu kandidat *biomarker* yang saat ini banyak diteliti adalah *long non-coding RNA* (lncRNA). Berbeda dengan anggapan sebelumnya yang menganggap lncRNA hanya sebagai sampah biologis, penelitian terkini menunjukkan bahwa lncRNA mengandung informasi genetik yang sangat penting dan berperan besar dalam proses regulasi gen. Dengan panjang lebih dari 200 nukleotida, lncRNA memiliki potensi tinggi sebagai *biomarker* karena menunjukkan pola ekspresi yang sangat spesifik pada jaringan, tipe sel, maupun tahapan perkembangan penyakit tertentu [6, 7].

Ketika terjadi kelainan pada lncRNA, proses regulasi gen dapat terganggu sehingga memicu perkembangan kanker. Kelainan tersebut dapat berupa perubahan jumlah ekspresi, perubahan fungsi biologis, maupun perubahan struktur lncRNA itu sendiri. Pada tingkat epigenetik, lncRNA abnormal dapat menghambat ataupun membuka akses terhadap DNA secara tidak normal sehingga menyebabkan gen penekan tumor (*tumor suppressor genes*) mengalami metilasi dan menjadi tidak aktif, sementara onkogen justru diaktifkan. Pada tingkat transkripsi, lncRNA abnormal dapat berikatan dengan faktor transkripsi sehingga mengubah fungsi normalnya dan menyebabkan produksi protein yang berbahaya secara berlebihan. Sementara itu, pada tingkat pascatranskripsi, lncRNA abnormal dapat bertindak sebagai *miRNA sponge*, yaitu mengikat miRNA sehingga kehilangan fungsinya. Akibatnya, onkogen dapat menghasilkan protein penyebab kanker tanpa adanya mekanisme penghambatan [8, 9].

Berbagai penelitian telah membuktikan bahwa lncRNA memiliki potensi besar sebagai *biomarker* kanker. lncRNA yang diekstraksi dari sampel darah pasien mampu memberikan tingkat akurasi deteksi yang lebih tinggi dibandingkan *biomarker* berbasis protein konvensional [10]. Selain itu, lncRNA juga telah berhasil dimanfaatkan untuk mendeteksi berbagai jenis kanker, termasuk kanker lambung, kanker payudara, kanker kolorektal, dan kanker paru-paru [11–14].

Meskipun lncRNA memiliki potensi yang sangat besar sebagai *biomarker*, analisis terhadap lncRNA bukanlah pekerjaan yang sederhana. Kompleksitas data biologis lncRNA, yang meliputi informasi *sequence*, struktur sekunder, tingkat ekspresi, serta interaksinya dengan berbagai molekul lain, menuntut penggunaan pendekatan komputasi yang mampu mengolah informasi tersebut secara efektif. Oleh karena itu, metode *machine learning* dan *deep learning* semakin banyak diterapkan dalam penelitian lncRNA. Berbagai algoritma pembelajaran mesin, seperti *Support Vector Machine* (SVM), *Random Forest*, dan *Extreme Gradient Boosting* (XGBoost), telah dimanfaatkan untuk melakukan klasifikasi lncRNA berdasarkan informasi *sequence* maupun struktur sekundernya [15, 16]. Selain itu, lncRNA juga sering dikombinasikan dengan molekul biologis lainnya, seperti *messenger RNA* (mRNA) dan *microRNA* (miRNA), untuk melakukan klasifikasi jaringan maupun identifikasi berbagai jenis kanker [17, 18]. Pendekatan tersebut telah berhasil diterapkan, antara lain, pada deteksi kanker usus besar dan kanker payudara [19, 20].

Di sisi lain, hubungan biologis antara lncRNA dan *RNA-binding protein* (RBP) kini semakin banyak direpresentasikan dalam bentuk graf sehingga melahirkan penerapan *Graph Neural Network* (GNN) pada bidang bioinformatika. Representasi berbasis graf dianggap lebih efektif karena mampu menggambarkan hubungan nyata antarmolekul secara langsung, sehingga informasi mengenai keterkaitan biologis dapat dipertahankan selama proses pembelajaran. Dalam penelitian prediksi hubungan miRNA dan penyakit, *Relational Graph Convolutional Networks* (RGCN) digunakan untuk melakukan klasifikasi pada graf heterogen miRNA-disease [21]. Pada penelitian lainnya, *Heterogeneous Graph Transformer* (HGT) digunakan untuk mengklasifikasikan hubungan antara circRNA dan obat [22]. Walaupun berbagai pendekatan tersebut telah menunjukkan hasil yang menjanjikan, penelitian yang secara khusus mengintegrasikan fitur struktur sekunder lncRNA, profil ekspresi lncRNA, serta hubungan antara lncRNA dengan *RNA-binding protein* (RBP) untuk mendeteksi kanker hati masih sangat terbatas. Selain itu, penggunaan metode *encoder* untuk menghasilkan *embedding* dari struktur sekunder lncRNA juga masih jarang ditemukan dalam penelitian sebelumnya.

Berdasarkan kondisi tersebut, terdapat dua permasalahan utama yang perlu diperhatikan. Pertama, representasi fitur lncRNA yang hanya bergantung pada informasi ekspresi belum sepenuhnya menggambarkan karakteristik molekuler lncRNA, karena struktur sekunder juga dapat memberikan informasi tambahan mengenai karakteristik molekul tersebut. Kedua, penggunaan informasi biologis secara terpisah dapat menyebabkan hubungan antar-entitas biologis tidak dimanfaatkan secara optimal. Dengan demikian, diperlukan suatu kerangka yang mampu melakukan pembelajaran representasi pada tingkat fitur sekaligus mempertahankan hubungan biologis pada tingkat graf.

Untuk mengatasi permasalahan tersebut, penelitian ini mengusulkan *framework Variational Autoencoder* (VAE)-GraphSAGE untuk klasifikasi lncRNA yang berasosiasi dengan kanker hati. VAE digunakan sebagai metode *representation learning* untuk menghasilkan representasi laten dari fitur yang berasal dari struktur sekunder lncRNA. Representasi laten tersebut kemudian dikombinasikan dengan profil ekspresi lncRNA untuk membentuk representasi fitur *node* yang lebih komprehensif. Selanjutnya, informasi tersebut digunakan untuk membangun

graf biologis heterogen yang tidak hanya merepresentasikan kemiripan antar-lncRNA, tetapi juga hubungan lncRNA-RBP dan interaksi *protein-protein* (PPI). GraphSAGE kemudian digunakan untuk melakukan klasifikasi *node* secara semi-induktif pada graf tersebut.

Kebaruhan utama penelitian ini terletak pada integrasi informasi biologis pada dua tingkat, yaitu tingkat fitur dan tingkat relasi. Pada tingkat fitur, representasi laten struktur sekunder yang diperoleh melalui VAE digabungkan dengan profil ekspresi lncRNA. Pada tingkat relasi, informasi similaritas antar-lncRNA, interaksi lncRNA-RBP, dan jaringan PPI diintegrasikan ke dalam *heterogeneous biological graph*. Dengan demikian, *framework* yang diusulkan tidak hanya memanfaatkan karakteristik individual lncRNA, tetapi juga memanfaatkan konteks hubungan biologisnya dalam proses klasifikasi. Selain itu, penelitian ini membandingkan VAE dengan beberapa metode representasi alternatif, yaitu *Autoencoder* (AE), *Multi-Layer Perceptron* (MLP), *Convolutional Neural Network* (CNN), dan *Residual Network* (ResNet), dengan GraphSAGE yang digunakan sebagai *classifier* yang sama. Perbandingan tersebut dilakukan untuk mengevaluasi kontribusi metode *representation learning* terhadap kualitas klasifikasi lncRNA. Selanjutnya, penelitian ini juga membandingkan hasil *embedding* VAE terhadap beberapa model klasifikasi selain GraphSAGE, yaitu HGT dan RGCN. Hal ini dilakukan untuk mengevaluasi bahwa perbedaan karakteristik model klasifikasi juga memengaruhi hasil klasifikasi.

Berdasarkan uraian tersebut, penelitian ini bertujuan mengembangkan *framework* VAE-GraphSAGE untuk mengintegrasikan informasi struktur sekunder, ekspresi, serta hubungan biologis lncRNA dalam graf heterogen untuk mengidentifikasi lncRNA yang berasosiasi dengan kanker hati. Penelitian ini juga bertujuan mengevaluasi efektivitas representasi berbasis VAE dibandingkan dengan metode representasi lainnya dan juga model klasifikasi lainnya dalam meningkatkan kemampuan klasifikasi. *Framework* yang diusulkan diharapkan dapat menghasilkan representasi lncRNA yang lebih informatif serta mendukung identifikasi kandidat lncRNA yang berpotensi digunakan sebagai *biomarker* kanker hati.

2. Metode

2.1. Dataset

Penelitian ini memanfaatkan beberapa jenis data yang diperoleh dari berbagai basis data biologis publik. Data utama yang digunakan meliputi data lncRNA pada jaringan hati normal maupun kanker, profil ekspresi lncRNA, struktur sekunder lncRNA, *RNA-binding protein* (RBP), serta interaksi antarprotein (*Protein-Protein Interaction*/PPI). Data lncRNA yang diketahui berasosiasi dengan kanker hati diperoleh dari basis data Lnc2Cancer 3.0 [23] dan LncRNADisease v3.0 [24]. Informasi mengenai *sequence*, posisi genomik, serta kromosom masing-masing lncRNA diperoleh dari GENCODE [25], yaitu basis data anotasi gen yang komprehensif namun tidak menyediakan informasi label kelas.

Oleh karena itu, untuk memperoleh kelas negatif, seluruh identitas lncRNA yang berasal dari Lnc2Cancer dan LncRNADisease dicocokkan dengan data pada GENCODE. Seluruh lncRNA yang terdapat pada GENCODE tetapi tidak ditemukan pada kedua basis data tersebut kemudian ditetapkan sebagai kelas negatif. Data profil ekspresi lncRNA pada jaringan kanker hati diperoleh dari *The Cancer Genome Atlas* (TCGA) [26], sedangkan data ekspresi pada jaringan hati normal diperoleh dari *Genotype-Tissue Expression* (GTEx) [27]. Struktur sekunder setiap lncRNA diperoleh melalui proses prediksi menggunakan perangkat lunak ViennaRNA RNAfold. Data *RNA-binding protein* (RBP) yang aktif pada jaringan hati diperoleh dari eCLIP ENCODE HepG2 [28]. Informasi mengenai *Protein Family* (Pfam) diperoleh dari UniProt [29], sedangkan data *Protein-Protein Interaction* (PPI) diperoleh dari IntAct [30]. Secara keseluruhan, dataset yang digunakan terdiri atas 1.000 lncRNA, yang meliputi 400 sampel positif (berasosiasi dengan

kanker hati) dan 600 sampel negatif. Selain itu, penelitian ini juga menggunakan 85 protein RBP sebagai *node* protein pada graf heterogen. Dataset lncRNA kemudian dibagi menjadi tiga *subset*, yaitu data *train*, data validasi, dan data *test* dengan rasio pembagian 70:15:15.

2.2. Integrasi VAE dan GraphSAGE

Penelitian ini mengusulkan suatu kerangka yang mengintegrasikan VAE dan GraphSAGE dalam dua tahap pembelajaran. Pada tahap pertama, VAE digunakan untuk mempelajari representasi laten struktur sekunder lncRNA melalui proses rekonstruksi tanpa menggunakan label kelas. Representasi laten yang dihasilkan kemudian digunakan sebagai bagian dari fitur *node* pada graf heterogen. Pada tahap kedua, GraphSAGE menggunakan fitur *node* dan hubungan biologis pada graf tersebut untuk melakukan klasifikasi lncRNA. Subbab ini menjabarkan landasan matematis kedua komponen tersebut sekaligus bagaimana keduanya disatukan menjadi satu alur komputasi.

VAE sebagai encoder struktur sekunder. VAE merupakan perluasan dari *Autoencoder* (AE). Berbeda dari AE konvensional yang memetakan *input* ke satu titik tetap di ruang laten, VAE memetakan setiap *input* ke distribusi probabilistik, yang umumnya diasumsikan berdistribusi Gaussian dengan parameter tertentu. Pendekatan ini memungkinkan model menangkap ketidakpastian dalam representasi laten, sekaligus menjaga ruang laten tetap terstruktur dan kontinu [31].

Proses pelatihan VAE dilakukan secara khusus untuk tujuan rekonstruksi struktur sekunder lncRNA. Pada setiap *epoch*, data *input* $x = \{x_1, x_2, \dots, x_n\}$ diproses oleh VAE untuk menghasilkan rekonstruksi $\hat{x}$, μ , dan $\log \sigma^2$. *Reconstruction loss* dihitung menggunakan *Mean Squared Error* (MSE),

$$\mathcal{L}_{rec} = MSE(\hat{x}, x), \quad (1)$$

sedangkan KL *divergence* dihitung berdasarkan parameter distribusi laten sebagai

$$\mathcal{L}_{KL} = -\frac{1}{2} \text{mean} (1 + \log \sigma^2 - \mu^2 - \exp(\log \sigma^2)). \quad (2)$$

Kedua komponen tersebut kemudian dikombinasikan menjadi fungsi *loss* VAE,

$$\mathcal{L}_{VAE} = \mathcal{L}_{rec} + \beta \mathcal{L}_{KL}. \quad (3)$$

Dengan β diberikan secara bertahap melalui mekanisme KL *annealing* hingga mencapai nilai maksimum 0,001. Komponen pertama, ekspektasi *log-likelihood* rekonstruksi, mendorong model merekonstruksi data seakurat mungkin dari sampel laten. Komponen kedua, KL *divergence*, mengukur seberapa jauh *posterior* q_ϕ menyimpang dari *prior* $p(z)$, dan bertindak sebagai regularisasi agar ruang laten tidak *collapse*. Karena proses *sampling* $z \sim q_\phi(z | x)$ bersifat stokastik dan tidak *differentiable*, *gradient* tidak dapat dipropagasi balik ke parameter *encoder* ϕ . Masalah ini diatasi dengan *reparameterization trick*:

$$z = \mu_\phi(x) + \sigma_\phi(x) \odot \epsilon, \quad \epsilon \sim \mathcal{N}(0, I), \quad (4)$$

sehingga sumber stokastisitas (ϵ) terpisah dari parameter yang dioptimasi (μ_ϕ, σ_ϕ), dan seluruh alur komputasi menjadi *differentiable* [31]. *Encoder* terlatih f_ϕ inilah yang menghasilkan *embedding* laten struktur sekunder $z_i^{VAE} = f_\phi(s_i) \in \mathbb{R}^d$ untuk setiap lncRNA i , dari data struktur sekunder s_i hasil prediksi ViennaRNA-RNAfold.

Fusi fitur node. *Embedding* z_i^{VAE} kemudian digabung dengan fitur ekspresi $e_i \in \mathbb{R}^9$ menjadi fitur *node* awal pada graf heterogen $G = (\mathcal{V}, \mathcal{E})$:

$$h_i^{(0)} = [z_i^{VAE}; e_i], \quad i \in \mathcal{V}_{lncRNA}, \quad (5)$$

sedangkan *node* RBP direpresentasikan sebagai vektor Pfam biner

$$h_j^{(0)} = p_j, \quad j \in \mathcal{V}_{RBP}.$$

Fusi ini didasarkan pada asumsi bahwa fitur ekspresi dan fitur struktur sekunder menangkap dua aspek biologis yang independen, ekspresi merepresentasikan seberapa aktif suatu lncRNA ditranskripsi, sedangkan struktur sekunder merepresentasikan struktur molekulnya sehingga menggabungkan keduanya menghasilkan ruang fitur yang lebih informatif dibanding menggunakan salah satunya saja.

GraphSAGE sebagai classifier pada graf. Fitur selanjutnya diproses oleh GraphSAGE, sebuah arsitektur *Graph Neural Network* (GNN) yang bersifat *inductive*. Dalam pembelajaran graf dikenal istilah *transductive* dan *inductive*. Pada pembelajaran *transductive*, *node test* tidak berlabel dan tidak dilibatkan dalam perhitungan *loss*, namun tetap dilibatkan dalam operasi *message passing*, sedangkan pada pembelajaran *inductive*, *node test* sama sekali tidak diamati selama *training*, baik dalam perhitungan *loss* maupun *message passing* [32]. GraphSAGE termasuk kategori *inductive*. Alih-alih menggunakan seluruh tetangga suatu *node*, GraphSAGE mengambil sampel sebagian tetangga secara acak dan mengagregasikan fiturnya, sehingga mampu menghasilkan *embedding* untuk *node* baru yang tidak terlihat saat *training* [33], menjadikannya cocok untuk aplikasi dinamis seperti prediksi fungsi lncRNA baru. Agregasi dalam GraphSAGE ada tiga, yaitu *Mean aggregator*, *LSTM aggregator*, dan *Pooling aggregator*, namun yang digunakan dalam penelitian ini adalah *mean aggregator*:

$$AGG_{mean} = MEAN \left(\{h_u^{k-1}, \forall u \in \mathcal{N}(v)\} \right). \quad (6)$$

Representasi *node* pada tiap *layer* $k = 1, 2, \dots, K$ diperbarui melalui:

$$h_v^k \leftarrow \sigma \left(W^k \cdot CONCAT \left(h_v^{k-1}, h_{\mathcal{N}(v)}^k \right) \right), \quad h_v^k \leftarrow \frac{h_v^k}{\|h_v^k\|_2}. \quad (7)$$

Menghasilkan *embedding* akhir $z_v = h_v^{(K)}$ untuk setiap $v \in \mathcal{V}$, dan prediksi kelas $\hat{y}_v = \text{softmax}(W \cdot z_v)$ untuk $v \in \mathcal{V}_{lncRNA}$. Model dilatih dengan meminimalkan *loss softmax* dan *negative log-likelihood*:

$$\mathcal{L}_{GraphSAGE}(\Theta) = \sum_{u \in \mathcal{V}_{train}} -\log(\text{softmax}(z_u, y_u)). \quad (8)$$

Skema pelatihan model hybrid. Karena z_i^{VAE} menjadi bagian dari *input* GraphSAGE, kedua model tidak dilatih bersamaan, melainkan diformulasikan sebagai optimisasi dua tahap sebagai berikut:

Tahap 1 (*embedding* fitur struktur sekunder):

$$\phi^* = \arg \min_{\theta, \phi} \mathcal{L}_{VAE}(\theta, \phi). \quad (9)$$

Tahap 2 (klasifikasi *node*):

$$\Theta^* = \arg \min_{\Theta} \mathcal{L}_{GraphSAGE}(\Theta; \phi^*), \quad (10)$$

dengan ϕ^* dibekukan pada tahap kedua.

2.3. Tahap Penelitian

Berikut penjabaran mengenai tahap-tahap yang dilakukan dalam penelitian ini:

1. *Pengumpulan Data*. Tahap pertama dilakukan dengan mengumpulkan seluruh data lncRNA yang diketahui berasosiasi dengan kanker hati dari berbagai sumber yang telah ditentukan. Setelah daftar nama lncRNA diperoleh, data lncRNA diambil dari basis data GENCODE dengan melakukan penyaringan terhadap lncRNA yang diekspresikan pada jaringan hati. Hasil penyaringan tersebut menghasilkan seluruh lncRNA yang diekspresikan pada jaringan hati, baik yang berasosiasi dengan kanker maupun yang tidak. Selanjutnya, ID lncRNA positif dibandingkan dengan data GENCODE menggunakan operasi irisan (*intersection*). Seluruh lncRNA yang tidak termasuk ke dalam kelompok positif kemudian ditetapkan sebagai lncRNA negatif. GENCODE juga menyediakan informasi mengenai *sequence* nukleotida serta posisi genomik setiap lncRNA. Selanjutnya, seluruh *RNA-binding protein* (RBP) yang berasal dari data eCLIP ENCODE HepG2 dikumpulkan. Berdasarkan identitas protein tersebut, informasi mengenai *Protein Family* (Pfam) diperoleh dari UniProt, sedangkan informasi *Protein-Protein Interaction* (PPI) diperoleh dari IntAct.
2. *Preprocessing*. Setelah seluruh data berhasil dikumpulkan, dilakukan proses pembersihan data (*data cleaning*) dengan menghapus seluruh data yang tidak lengkap maupun memiliki nilai yang hilang (*missing values*). Tahap ini menghasilkan dataset akhir yang terdiri atas 400 lncRNA positif, 600 lncRNA negatif, dan 85 *RNA-binding protein* (RBP).
3. *Ekstraksi Fitur Node*.
 - Fitur struktur sekunder lncRNA diperoleh menggunakan model *Variational Autoencoder* (VAE) yang telah dilatih sebelumnya (*pre-trained*). Proses pelatihan VAE dilakukan terlebih dahulu agar representasi laten (*latent representation*) yang dihasilkan benar-benar mampu menangkap karakteristik biologis struktur sekunder lncRNA, bukan sekadar menghasilkan representasi acak. Representasi laten yang dihasilkan oleh *encoder* kemudian digunakan sebagai fitur struktur sekunder bagi seluruh lncRNA pada data *train*, validasi, maupun *test*.
 - Profil ekspresi lncRNA direpresentasikan menggunakan vektor statistik deskriptif yang terdiri atas sembilan fitur, yaitu: rata-rata ekspresi pada data GTEx, rata-rata ekspresi pada data TCGA, varians ekspresi pada GTEx, varians ekspresi pada TCGA, median ekspresi pada GTEx, median ekspresi pada TCGA, proporsi transkrip aktif pada GTEx, proporsi transkrip aktif pada TCGA, nilai *log fold change*. Kesembilan nilai tersebut membentuk sebuah vektor berdimensi sembilan yang digunakan sebagai representasi profil ekspresi masing-masing lncRNA.
 - Fitur *Protein Family* (Pfam) diperoleh dari basis data UniProt dalam bentuk vektor biner yang menunjukkan keberadaan suatu keluarga protein pada masing-masing RBP. Karena setiap RBP hanya termasuk ke dalam satu atau beberapa keluarga protein tertentu dengan tingkat variasi yang cukup tinggi, maka hanya keluarga protein yang muncul sedikitnya satu kali pada keseluruhan himpunan RBP yang digunakan sebagai fitur.
4. *Ekstraksi Fitur Edge*.
 - Kemiripan antar-lncRNA dihitung menggunakan korelasi Pearson dengan mempertimbangkan dua jenis informasi, yaitu kemiripan profil ekspresi dan kemiripan struktur sekunder, yang dirumuskan sebagai berikut:

$$\mathcal{E}_{sim} = \alpha \cdot sim_{ss} + (1 - \alpha) \cdot sim_{eks}, \quad (11)$$

dengan sim_{ss} merupakan similaritas lncRNA berdasarkan struktur sekundernya, sim_{eks}

merupakan similaritas lncRNA berdasarkan ekspresinya, dan α merupakan bobot untuk masing-masing similaritas. Pada penelitian ini kedua komponen tersebut diberikan bobot yang sama, yaitu 50% untuk ekspresi dan 50% untuk struktur sekunder. Untuk mengurangi *noise* pada graf, setiap lncRNA hanya dihubungkan dengan 10 lncRNA yang memiliki nilai kemiripan tertinggi menggunakan pendekatan *Top-K*.

- Hubungan antara lncRNA dan *RNA-binding protein* dihitung berdasarkan posisi *binding site* masing-masing RBP pada setiap lncRNA dengan mempertimbangkan lokasi kromosom tempat interaksi tersebut terjadi. Nilai *binding* yang diperoleh selanjutnya digunakan sebagai bobot *edge* yang menghubungkan *node* lncRNA dengan *node* RBP.
- Hubungan antarprotein diperoleh dari basis data IntAct dalam bentuk skor interaksi protein-protein (PPI). Nilai tersebut digunakan sebagai bobot *edge* yang menghubungkan antar-*node* RBP pada graf heterogen.

5. *Klasifikasi*. Setelah seluruh fitur *node* dan fitur *edge* selesai dipersiapkan, seluruh informasi tersebut digunakan untuk membangun sebuah graf heterogen.

Graf ini terdiri atas *node* lncRNA, *node* RBP, *edge* similaritas antar-lncRNA, *edge binding* antara lncRNA dan RBP, dan *edge* PPI antar-RBP. Selanjutnya dilakukan proses *training* menggunakan arsitektur GraphSAGE. Pada penelitian ini, proses optimasi model difokuskan pada peningkatan nilai *precision*, sehingga model mampu mengidentifikasi lncRNA yang berasosiasi dengan kanker hati secara lebih akurat dan meminimalkan prediksi positif yang salah.

6. *Evaluasi Model*. Kinerja model dievaluasi menggunakan beberapa metrik klasifikasi yang umum digunakan, yaitu *Precision*, *Recall*, *F1-score*, dan *AUC score*. Keempat metrik tersebut digunakan untuk mengevaluasi kemampuan model secara menyeluruh dalam membedakan lncRNA yang berasosiasi maupun yang tidak berasosiasi dengan kanker hati.

3. Hasil dan Pembahasan

3.1. Keterbatasan Penelitian

Penelitian ini memiliki beberapa keterbatasan yang perlu diperhatikan dalam menginterpretasikan hasil. Pertama, ukuran dataset lncRNA yang digunakan relatif terbatas, yaitu sebanyak 1.000 lncRNA yang terdiri atas 400 lncRNA positif dan 600 lncRNA negatif, serta 85 RBP sebagai *node* protein pada graf heterogen. Meskipun pembagian data telah dilakukan menjadi data *training*, validasi, dan *testing*, jumlah tersebut masih relatif kecil dibandingkan dengan keseluruhan lncRNA yang terdapat pada genom manusia. Selain itu, data lncRNA yang digunakan dalam penelitian ini berasal dari sumber basis data publik yang tersedia dan berfokus pada konteks kanker hati, sehingga variasi biologis yang dapat direpresentasikan oleh dataset masih terbatas. Oleh karena itu, performa model yang diperoleh perlu diinterpretasikan dengan mempertimbangkan keterbatasan ukuran dan cakupan dataset, serta perlu divalidasi menggunakan dataset yang lebih besar dan independen.

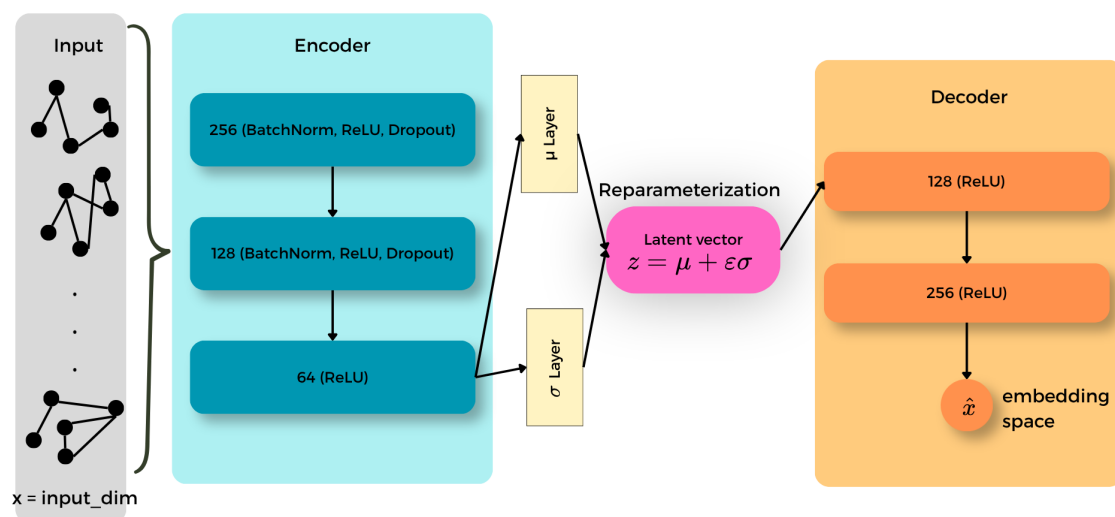
Kedua, penentuan kelas negatif masih memiliki potensi *label noise*. lncRNA yang digunakan sebagai kelas positif diperoleh dari Lnc2Cancer dan LncRNADisease, sedangkan lncRNA yang terdapat pada GENCODE tetapi tidak ditemukan pada kedua basis data tersebut ditetapkan sebagai kelas negatif. Strategi tersebut diperlukan karena GENCODE tidak menyediakan label asosiasi kanker, tetapi ketidaktercantuman suatu lncRNA pada Lnc2Cancer atau LncRNADisease tidak serta-merta menunjukkan bahwa lncRNA tersebut benar-benar tidak berasosiasi dengan kanker hati. lncRNA tersebut dapat saja belum terdokumentasi atau belum diteliti secara memadai. Dengan demikian, sebagian lncRNA yang diberi label negatif berpotensi merupakan lncRNA yang sebenarnya memiliki hubungan dengan kanker hati tetapi belum tercatat

dalam basis data yang digunakan.

Ketiga, terdapat potensi *batch effect* pada data ekspresi karena data diperoleh dari dua sumber independen, yaitu TCGA untuk jaringan kanker hati dan GTEx untuk jaringan hati normal. Setiap lncRNA dalam dataset memiliki informasi ekspresi dari kedua sumber tersebut, yang kemudian dirangkum menjadi fitur statistik berupa rata-rata, varians, median, proporsi transkrip aktif, dan *log fold change*. Meskipun penggunaan kedua sumber memungkinkan karakterisasi perbedaan ekspresi antara kondisi kanker dan normal, perbedaan prosedur pengambilan sampel, karakteristik sampel, dan pemrosesan data antara TCGA dan GTEx berpotensi menghasilkan variasi teknis yang tidak sepenuhnya berkaitan dengan kondisi biologis. Selain itu, karena TCGA secara langsung merepresentasikan kondisi kanker sedangkan GTEx merepresentasikan kondisi normal, pengaruh sumber data dan kondisi biologis tidak dapat sepenuhnya dipisahkan dalam penelitian ini.

3.2. Embedding Struktur Sekunder lncRNA

Eksperimen utama dalam penelitian ini berfokus pada penerapan metode *encoder* untuk menghasilkan representasi (*embedding*) dari struktur sekunder lncRNA. Model utama yang digunakan adalah *Variational Autoencoder* (VAE) dengan arsitektur sebagaimana ditunjukkan pada Gambar 1.



Gambar 1. Arsitektur VAE.

Model VAE terlebih dahulu di-*train* menggunakan data struktur sekunder lncRNA. Setelah proses *training* selesai, performa model dievaluasi menggunakan data validasi. Selanjutnya, parameter model dibekukan (*freeze*) sehingga *encoder* dapat digunakan untuk menghasilkan representasi laten pada data pelatihan, validasi, maupun data *test*.

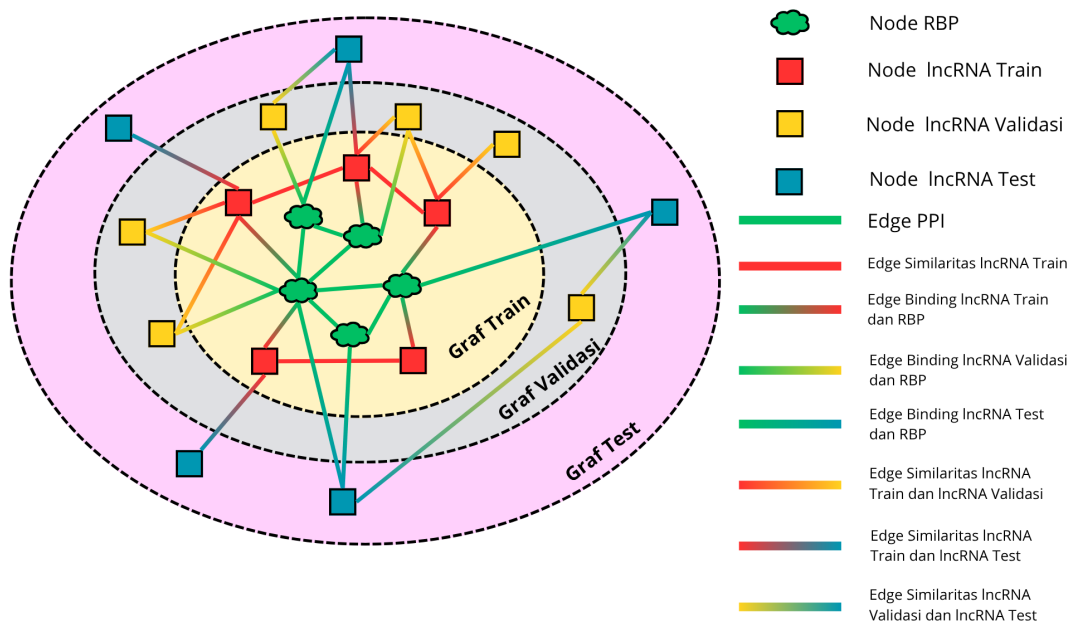
Model VAE mempelajari struktur sekunder lncRNA dan dievaluasi menggunakan data validasi, selanjutnya, parameter-parameternya dibekukan dan digunakan untuk melakukan *embedding* terhadap data *train*, validasi, dan *test*. *Input* dari setiap lncRNA memiliki 9 dimensi, yang akan di-*embed* menggunakan VAE, terdiri dari persentase kemunculan keempat basa sepanjang urutan, kandungan GC, panjang *sequence*, nilai rata-rata, nilai maksimum, dan simpangan baku nukleotida yang tidak berpasangan.

Kolom GC mewakili rasio kandungan GC terhadap panjang *sequence*. Norm panjang mengacu pada normalisasi panjang urutan menjadi 5.000 nukleotida. Frekuensi A, U, G, dan C masing-masing mewakili persentase basa adenin, urasil, guanin, dan sitosin dalam *sequence*.

Rata-rata *unpaired* adalah nilai rata-rata probabilitas basa yang tidak berpasangan di seluruh urutan, yang menunjukkan seberapa terbuka molekul RNA secara keseluruhan. Maksimum *unpaired* adalah nilai probabilitas tertinggi, yang menunjukkan area yang paling mudah diakses oleh molekul lain. Standar deviasi *unpaired* menunjukkan tingkat variasi dalam perbedaan antara daerah terbuka dan tertutup (berpasangan) dalam RNA.

3.3. Hasil Klasifikasi

Untuk klasifikasi, hanya lncRNA yang memiliki kelas. Konstruksi graf untuk proses klasifikasi diberikan pada Gambar 2.

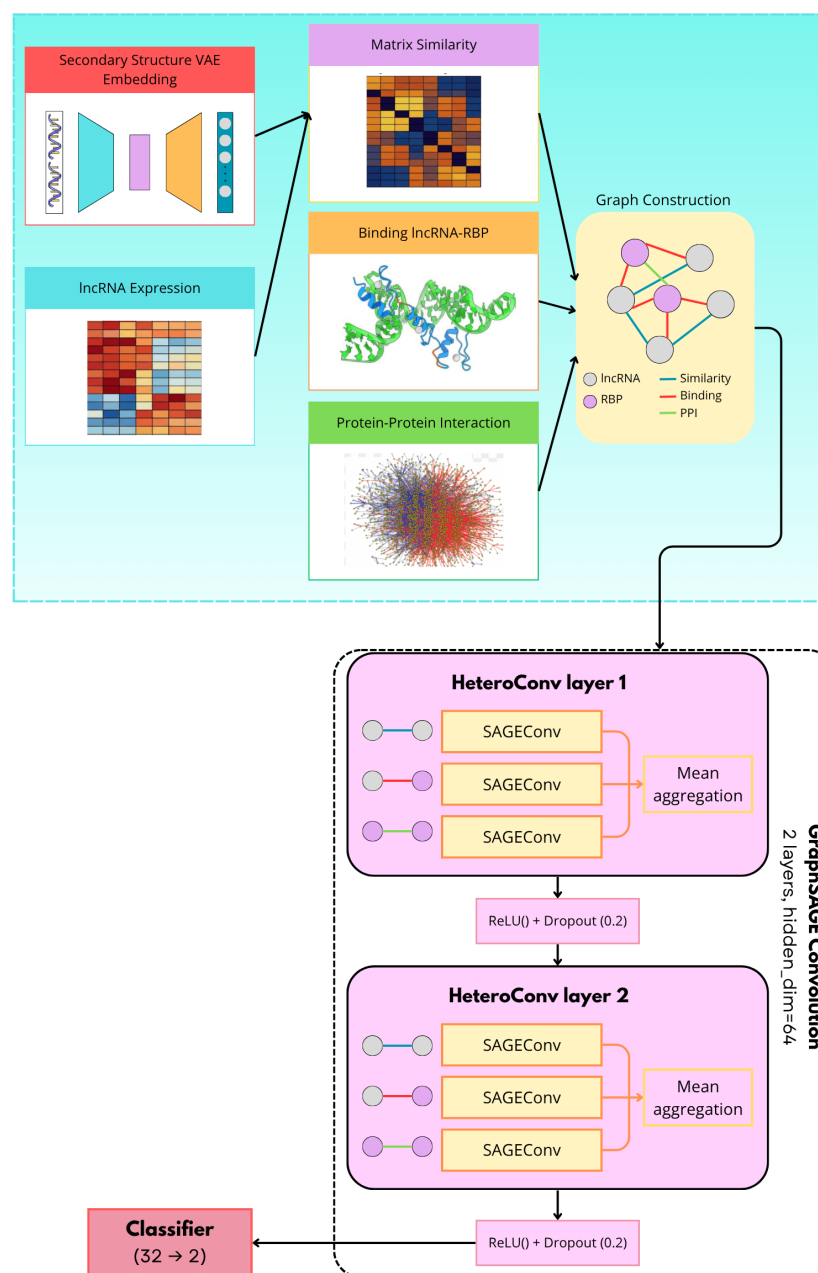


Gambar 2. Konstruksi Graf Train, Validasi, dan Test.

Pembuatan graf pada penelitian ini diawali dengan membagi *node* lncRNA menjadi tiga kelompok, yaitu data *train*, validasi, dan *test*, yang masing-masing ditunjukkan dengan warna merah, kuning, dan biru. Sementara itu, RBP digunakan sebagai *node* yang ditunjukkan dengan warna hijau. Graf dibangun sebagai graf heterogen karena terdapat dua tipe *node*, yaitu lncRNA dan RBP, serta beberapa tipe hubungan. Hubungan antar-RBP direpresentasikan sebagai PPI (*Protein-Protein Interaction*), sedangkan hubungan antar-lncRNA dibentuk berdasarkan kemiripan struktur dan ekspresi lncRNA. Selain itu, hubungan antara lncRNA dan RBP dibentuk berdasarkan informasi *binding site* yang diperoleh dari data interaksi lncRNA-RBP. Dengan demikian, struktur graf tidak hanya merepresentasikan kemiripan antar-lncRNA, tetapi juga memasukkan informasi biologis berupa interaksi RBP.

Pada graf *train*, *node* lncRNA *train* digunakan sebagai *node* utama untuk membentuk hubungan kemiripan antar-lncRNA yang ditunjukkan oleh garis merah, sedangkan hubungan lncRNA dengan RBP ditunjukkan melalui garis dengan warna hijau-merah. RBP yang digunakan pada graf *train* juga terhubung satu sama lain melalui *edge* PPI berwarna hijau. Selanjutnya, data validasi ditambahkan untuk mengevaluasi kemampuan model terhadap *node* yang tidak digunakan dalam proses pembelajaran parameter model. Hubungan *binding* antara lncRNA validasi dan RBP ditunjukkan dengan garis hijau-kuning, sedangkan hubungan kemiripan antara lncRNA *train* dan validasi ditunjukkan dengan garis merah-kuning. Dengan cara ini, informasi struktural graf dari data *train* tetap dapat digunakan untuk menghubungkan *node* validasi tanpa menjadikan label validasi sebagai bagian dari proses pelatihan.

Demikian pula, data *test* ditempatkan sebagai kelompok *node* yang digunakan pada tahap evaluasi akhir dan ditunjukkan dengan warna biru. Hubungan antara lncRNA *test* dan RBP di-representasikan dengan garis hijau–biru, sedangkan hubungan kemiripan antara lncRNA *train* dan *test* ditunjukkan dengan garis merah–biru. Pemisahan tersebut bertujuan menjaga agar label dari data validasi dan *test* tidak digunakan dalam proses pembelajaran model. Secara keseluruhan, graf menggambarkan bahwa informasi biologis dari RBP dan hubungan kemiripan lncRNA dimanfaatkan untuk membangun konektivitas antar-*node* pada masing-masing tahap data, sementara label kelas hanya digunakan pada data *train* selama proses pembelajaran dan data validasi serta *test* digunakan untuk mengukur kemampuan generalisasi model. Klasifikasi lncRNA dilakukan menggunakan GraphSAGE. Secara keseluruhan, alur kerja yang mengarah ke tahap klasifikasi diberikan pada **Gambar 3**.



Gambar 3. Arsitektur GraphSAGE.

Arsitektur model dimulai dari tahap ekstraksi dan integrasi informasi biologis lncRNA sebelum dilakukan proses klasifikasi. Pada tahap awal, informasi struktur sekunder lncRNA

diproses menggunakan *Variational Autoencoder* (VAE) untuk menghasilkan representasi atau *embedding* berdimensi lebih rendah. Secara paralel, data ekspresi lncRNA digunakan sebagai representasi karakteristik ekspresi masing-masing lncRNA. Kedua informasi tersebut kemudian digunakan dalam pembentukan matriks kemiripan (*similarity matrix*) antar-lncRNA. Selain informasi kemiripan, graf juga dibangun dengan memanfaatkan dua jenis interaksi biologis, yaitu *binding* lncRNA–RBP yang merepresentasikan hubungan antara lncRNA dan *RNA-binding protein* serta *protein–protein interaction* (PPI) yang merepresentasikan hubungan antar-RBP. Ketiga sumber informasi tersebut, yaitu kemiripan lncRNA, *binding* lncRNA–RBP, dan PPI, selanjutnya diintegrasikan untuk membentuk *heterogeneous graph* yang terdiri atas dua tipe *node*, yaitu lncRNA dan RBP, serta beberapa tipe *edge*.

Graf heterogen yang telah terbentuk kemudian menjadi masukan bagi arsitektur GraphSAGE melalui dua lapisan HeteroConv. Pada HeteroConv *layer* 1, setiap jenis relasi diproses menggunakan SAGEConv secara terpisah sehingga informasi dari relasi kemiripan lncRNA, *binding* lncRNA–RBP, dan PPI dapat dipelajari sesuai karakteristik masing-masing hubungan. Hasil dari masing-masing SAGEConv kemudian digabungkan menggunakan *mean aggregation* untuk memperoleh representasi *node* yang telah mengintegrasikan informasi dari berbagai relasi. Hasil agregasi selanjutnya melewati fungsi aktivasi ReLU dan *dropout* sebesar 0,2 untuk memperkenalkan non-linearitas sekaligus mengurangi risiko *overfitting*. Representasi hasil *layer* pertama kemudian diteruskan ke HeteroConv *layer* 2 dengan mekanisme yang sama, yaitu pemrosesan setiap relasi menggunakan SAGEConv, diikuti oleh *mean aggregation*. Setelah itu, representasi kembali diproses menggunakan ReLU dan *dropout* 0,2.

Dengan demikian, urutan pemrosesan model sebelum klasifikasi adalah ekstraksi *embedding* struktur sekunder menggunakan VAE → pemanfaatan informasi ekspresi → pembentukan matriks *similarity* → integrasi informasi *binding* lncRNA–RBP dan PPI → konstruksi *heterogeneous graph* → HeteroConv *layer* 1 → ReLU dan *dropout* → HeteroConv *layer* 2 → ReLU dan *dropout*. Representasi *node* yang dihasilkan setelah dua lapisan GraphSAGE tersebut kemudian menjadi representasi akhir yang digunakan oleh *classifier* dengan struktur $32 \rightarrow 2$, yang selanjutnya menghasilkan prediksi kelas lncRNA. Arsitektur ini memungkinkan informasi lokal dari berbagai jenis hubungan biologis dikombinasikan secara bertahap sebelum representasi akhir digunakan untuk menentukan kelas lncRNA.

3.4. Perbandingan Model

Dalam penelitian ini, model VAE-GraphSAGE dibandingkan dengan beberapa model lain, seperti AE, MLP, CNN, dan ResNet, sebagai tolok ukur untuk VAE dengan menggunakan 5 buah *seed*, dan disajikan pada Tabel 1.

Tabel 1. Perbandingan performa model *embedding*.

	VAE-GraphSAGE	AE-GraphSAGE	ResNet-GraphSAGE	MLP-GraphSAGE	CNN-GraphSAGE
<i>Accuracy</i>	0,8813 ± 0,0288	0,8733 ± 0,0236	0,8853 ± 0,0251	0,9093 ± 0,0060	0,8733 ± 0,0302
<i>Precision</i>	0,8356 ± 0,0640	0,8540 ± 0,0607	0,8715 ± 0,0747	0,9472 ± 0,0288	0,8334 ± 0,0722
<i>Recall</i>	0,8833 ± 0,0204	0,8300 ± 0,0183	0,8467 ± 0,0380	0,8200 ± 0,0139	0,8667 ± 0,0425
<i>Specificity</i>	0,8800 ± 0,0579	0,9022 ± 0,0474	0,9111 ± 0,0619	0,9689 ± 0,0183	0,8778 ± 0,0724
<i>F1-score</i>	0,8573 ± 0,0281	0,8405 ± 0,0234	0,8562 ± 0,0238	0,8787 ± 0,0056	0,8467 ± 0,0252
ROC-AUC	0,9495 ± 0,0036	0,9477 ± 0,0030	0,9469 ± 0,0045	0,9501 ± 0,0042	0,9577 ± 0,0019
TP	53,0 ± 1,2247	49,8 ± 1,0954	50,8 ± 2,2804	49,2 ± 0,8367	52,0 ± 2,5495
TN	79,2 ± 5,2154	81,2 ± 4,2661	82,0 ± 5,5678	87,2 ± 1,6432	79,0 ± 6,5192
FP	10,8 ± 5,2154	8,8 ± 4,2661	8,0 ± 5,5678	2,8 ± 1,6432	11,0 ± 6,5192
FN	7,0 ± 1,2247	10,2 ± 1,0954	9,2 ± 2,2804	10,8 ± 0,8367	8,0 ± 2,5495

Berdasarkan Tabel 1, seluruh varian model menunjukkan kinerja klasifikasi yang relatif baik dengan nilai ROC-AUC berada pada rentang 0,9469–0,9577. Model MLP-GraphSAGE memperoleh nilai *accuracy* tertinggi sebesar $0,9093 \pm 0,0060$, diikuti oleh ResNet-GraphSAGE sebesar $0,8853 \pm 0,0251$, VAE-GraphSAGE sebesar $0,8813 \pm 0,0288$, serta AE-GraphSAGE dan CNN-GraphSAGE yang sama-sama sebesar 0,8733, masing-masing dengan standar deviasi 0,0236 dan 0,0302. Meskipun demikian, perbedaan *accuracy* tidak secara langsung menunjukkan bahwa MLP-GraphSAGE merupakan model terbaik secara keseluruhan, karena metrik lainnya perlu dipertimbangkan, terutama pada data klasifikasi yang melibatkan kelas positif dan negatif.

Pada metrik *precision*, *recall*, dan *F1-score*, terlihat karakteristik kinerja yang berbeda pada masing-masing model. MLP-GraphSAGE menghasilkan *precision* tertinggi sebesar $0,9472 \pm 0,0288$ dan *F1-score* tertinggi sebesar $0,8787 \pm 0,0056$, yang menunjukkan kemampuan paling baik dalam menghasilkan prediksi positif yang tepat sekaligus mempertahankan keseimbangan antara *precision* dan *recall*. Sebaliknya, VAE-GraphSAGE menghasilkan *recall* tertinggi sebesar $0,8833 \pm 0,0204$, lebih tinggi dibandingkan CNN-GraphSAGE sebesar $0,8667 \pm 0,0425$ dan ResNet-GraphSAGE sebesar $0,8467 \pm 0,0380$. Hasil tersebut menunjukkan bahwa VAE-GraphSAGE memiliki kemampuan yang lebih baik dalam mendeteksi sampel positif, sehingga lebih sedikit kasus positif yang terlewat. Hal ini juga tercermin dari nilai FN VAE-GraphSAGE yang paling rendah, yaitu $7,0 \pm 1,2247$.

Pada metrik *specificity*, MLP-GraphSAGE kembali menunjukkan kinerja paling tinggi dengan nilai $0,9689 \pm 0,0183$, yang menunjukkan kemampuan paling baik dalam mengidentifikasi sampel negatif secara benar. Nilai ini didukung oleh jumlah FP yang paling rendah, yaitu $2,8 \pm 1,6432$. Sementara itu, VAE-GraphSAGE memiliki *specificity* sebesar $0,8800 \pm 0,0579$ dengan FP sebesar $10,8 \pm 5,2154$. Menariknya, meskipun VAE-GraphSAGE tidak unggul pada *specificity*, model tersebut memiliki jumlah TP tertinggi sebesar $53,0 \pm 1,2247$. Dengan demikian, VAE-GraphSAGE cenderung lebih sensitif dalam mengenali kelas positif, sedangkan MLP-GraphSAGE lebih konservatif dan lebih efektif dalam menekan kesalahan prediksi positif maupun negatif.

Dari sisi kemampuan diskriminasi secara keseluruhan, CNN-GraphSAGE menghasilkan ROC-AUC tertinggi sebesar $0,9577 \pm 0,0019$, sedikit lebih tinggi dibandingkan MLP-GraphSAGE sebesar $0,9501 \pm 0,0042$ dan VAE-GraphSAGE sebesar $0,9495 \pm 0,0036$. Namun, nilai ROC-AUC yang relatif berdekatan menunjukkan bahwa seluruh model memiliki kemampuan diskriminasi yang baik. Selain itu, VAE-GraphSAGE memiliki standar deviasi yang relatif kecil pada *recall*, *F1-score*, dan ROC-AUC, yang menunjukkan performa yang cukup konsisten antarpercobaan. Secara keseluruhan, MLP-GraphSAGE memberikan performa terbaik berdasarkan *accuracy*, *precision*, *specificity*, dan *F1-score*, sedangkan CNN-GraphSAGE unggul pada ROC-AUC dan VAE-GraphSAGE unggul pada *recall*. Oleh karena itu, pemilihan model tidak hanya didasarkan pada satu metrik, tetapi perlu disesuaikan dengan tujuan klasifikasi. Karena prioritas utama penelitian ini adalah mendeteksi sebanyak mungkin sampel positif kanker, tingginya *recall* VAE-GraphSAGE menjadi karakteristik yang penting untuk dipertimbangkan.

Selanjutnya, sebagai pertimbangan untuk menentukan model *classifier*, diberikan perbandingan performa integrasi *embedding* VAE dengan beberapa model *classifier* berbeda, yaitu GraphSAGE, HGT, dan RGCN, pada Tabel 2.

Berdasarkan Tabel 2, perbandingan dilakukan dengan menggunakan parameter dan representasi fitur yang sama, yaitu fitur hasil ekstraksi *Variational Autoencoder* (VAE), sehingga perbedaan performa terutama mencerminkan pengaruh model *Graph Neural Network* (GNN) yang digunakan sebagai *classifier*. VAE-GraphSAGE menghasilkan *accuracy* tertinggi sebesar $0,8813 \pm 0,0288$, diikuti oleh VAE-HGT sebesar $0,8627 \pm 0,0243$ dan VAE-RGCN sebesar $0,8427 \pm$

Tabel 2. Perbandingan model *classifier*.

	VAE-GraphSAGE	VAE-HGT	VAE-RGCN
<i>Accuracy</i>	$0,8813 \pm 0,0288$	$0,8627 \pm 0,0243$	$0,8427 \pm 0,0285$
<i>Precision</i>	$0,8356 \pm 0,0640$	$0,8153 \pm 0,0345$	$0,7996 \pm 0,0584$
<i>Recall</i>	$0,8833 \pm 0,0204$	$0,8500 \pm 0,0312$	$0,8167 \pm 0,0408$
<i>Specificity</i>	$0,8800 \pm 0,0579$	$0,8711 \pm 0,0268$	$0,8600 \pm 0,0536$
<i>F1-score</i>	$0,8573 \pm 0,0281$	$0,8320 \pm 0,0293$	$0,8064 \pm 0,0302$
ROC-AUC	$0,9495 \pm 0,0036$	$0,9425 \pm 0,0083$	$0,9283 \pm 0,0106$
TP	$53,0 \pm 1,2247$	$51,0 \pm 1,8708$	$49,0 \pm 2,4495$
TN	$79,2 \pm 5,2154$	$78,4 \pm 2,4083$	$77,4 \pm 4,8270$
FP	$10,8 \pm 5,2154$	$11,6 \pm 2,4083$	$12,6 \pm 4,8270$
FN	$7,0 \pm 1,2247$	$9,0 \pm 1,8708$	$11,0 \pm 2,4495$

0,0285. Pola yang sama juga terlihat pada *F1-score*, dengan VAE-GraphSAGE memperoleh nilai $0,8573 \pm 0,0281$, lebih tinggi dibandingkan VAE-HGT sebesar $0,8320 \pm 0,0293$ dan VAE-RGCN sebesar $0,8064 \pm 0,0302$. Hasil tersebut menunjukkan bahwa ketika fitur masukan dibuat tetap, GraphSAGE mampu memanfaatkan informasi pada struktur graf secara lebih efektif dibandingkan HGT dan RGCN untuk tugas klasifikasi yang dilakukan.

Pada *precision* dan *recall*, VAE-GraphSAGE juga menunjukkan performa terbaik. Nilai *precision* VAE-GraphSAGE mencapai $0,8356 \pm 0,0640$, sedangkan VAE-HGT dan VAE-RGCN masing-masing memperoleh $0,8153 \pm 0,0345$ dan $0,7996 \pm 0,0584$. Sementara itu, nilai *recall* VAE-GraphSAGE sebesar $0,8833 \pm 0,0204$ lebih tinggi dibandingkan VAE-HGT sebesar $0,8500 \pm 0,0312$ dan VAE-RGCN sebesar $0,8167 \pm 0,0408$. Tingginya *recall* tersebut didukung oleh jumlah *true positive* (TP) yang paling besar pada VAE-GraphSAGE, yaitu $53,0 \pm 1,2247$, serta *false negative* (FN) yang paling rendah, yaitu $7,0 \pm 1,2247$. Dengan demikian, GraphSAGE menunjukkan kemampuan yang lebih baik dalam mempertahankan deteksi terhadap sampel positif dibandingkan kedua *classifier* lainnya.

Perbedaan performa juga terlihat pada kemampuan model dalam mengidentifikasi sampel negatif. VAE-GraphSAGE memperoleh *specificity* sebesar $0,8800 \pm 0,0579$, sedikit lebih tinggi dibandingkan VAE-HGT sebesar $0,8711 \pm 0,0268$ dan VAE-RGCN sebesar $0,8600 \pm 0,0536$. Hal ini sejalan dengan jumlah *false positive* (FP) VAE-GraphSAGE yang paling rendah, yaitu $10,8 \pm 5,2154$, dibandingkan VAE-HGT sebesar $11,6 \pm 2,4083$ dan VAE-RGCN sebesar $12,6 \pm 4,8270$. Dengan kata lain, GraphSAGE tidak hanya lebih baik dalam mengenali sampel positif, tetapi juga menghasilkan lebih sedikit kesalahan dalam mengklasifikasikan sampel negatif sebagai positif.

Pada metrik ROC-AUC, VAE-GraphSAGE kembali memperoleh nilai tertinggi sebesar $0,9495 \pm 0,0036$, dibandingkan VAE-HGT sebesar $0,9425 \pm 0,0083$ dan VAE-RGCN sebesar $0,9283 \pm 0,0106$. Selain memiliki nilai rata-rata tertinggi, standar deviasi ROC-AUC pada VAE-GraphSAGE juga paling kecil, yang menunjukkan performa diskriminasi yang relatif lebih konsisten antarpercobaan. Secara keseluruhan, hasil ini menunjukkan bahwa dengan representasi fitur VAE yang sama, GraphSAGE memberikan performa klasifikasi terbaik dibandingkan HGT dan RGCN pada seluruh metrik utama yang dievaluasi. Temuan ini mendukung penggunaan GraphSAGE sebagai *classifier* dalam model yang diusulkan karena mampu memanfaatkan hubungan antar-node pada graf secara efektif sekaligus mempertahankan keseimbangan antara kemampuan mendeteksi kelas positif dan mengidentifikasi kelas negatif.

4. Kesimpulan

Berdasarkan hasil pengujian, penelitian ini menunjukkan bahwa penggunaan VAE sebagai metode *embedding* struktur sekunder lncRNA yang dikombinasikan dengan GraphSAGE mampu menghasilkan performa klasifikasi yang kompetitif dalam mengidentifikasi lncRNA yang berasosiasi dengan kanker hati. Pada perbandingan metode *embedding* dengan GraphSAGE sebagai *classifier*, VAE-GraphSAGE memiliki keunggulan terutama pada *recall*, yang menunjukkan kemampuan lebih baik dalam mendeteksi lncRNA positif dan mengurangi sampel positif yang terlewat, sedangkan MLP-GraphSAGE memberikan performa terbaik pada *accuracy*, *precision*, *specificity*, dan *F1-score*, serta CNN-GraphSAGE menghasilkan ROC-AUC tertinggi. Selanjutnya, ketika representasi VAE yang sama digunakan pada beberapa *classifier*, VAE-GraphSAGE menghasilkan performa terbaik dibandingkan VAE-HGT dan VAE-RGCN pada metrik utama yang dievaluasi, menunjukkan bahwa GraphSAGE lebih efektif dalam memanfaatkan hubungan pada graf heterogen untuk tugas klasifikasi yang dilakukan. Hasil tersebut menunjukkan bahwa kontribusi VAE terletak pada kemampuannya menghasilkan representasi struktur sekunder yang mendukung deteksi kelas positif, sedangkan pemilihan *classifier* tetap memengaruhi karakteristik performa akhir model.

Kontribusi Penulis. Theodora Tantri Trisnawati: perancangan penelitian, pengumpulan dan analisis data, pengembangan model komputasi, interpretasi hasil penelitian, penyusunan gambar, serta penulisan naskah awal. Mohammad Isa Irawan: supervisi, memberi arahan konseptual, penelaahan metodologi dan hasil penelitian, revisi kritis terhadap naskah, serta persetujuan terhadap versi akhir naskah sebelum dipublikasikan.

Ucapan Terima Kasih. Penulis mengucapkan terima kasih kepada seluruh individu dan institusi yang telah memberikan bantuan dalam penyusunan penelitian ini. Penulis juga menyampaikan penghargaan kepada seluruh penyedia dan pengelola basis data biologis publik yang digunakan dalam penelitian ini. Upaya mereka dalam mengumpulkan, memelihara, dan menyediakan dataset berkualitas tinggi memberikan kontribusi yang sangat penting terhadap keberhasilan penelitian ini. Penulis juga mengucapkan terima kasih kepada *editor* serta para *reviewer* anonim atas berbagai masukan dan saran yang konstruktif sehingga kualitas serta kejelasan naskah ini dapat ditingkatkan secara signifikan.

Pembiayaan. Penelitian ini tidak menerima pendanaan eksternal.

Konflik Kepentingan. Para penulis menyatakan tidak ada konflik kepentingan yang terkait dengan artikel ini.

Ketersediaan Data. Data yang digunakan dalam penelitian ini menggunakan basis data publik, yaitu Lnc2Cancer [23], LncRNADisease [24], GENCODE [25], TCGA [26], GTEx [27], ENCODE [28], UniProt [29], dan IntAct [30].

Referensi

- [1] International Agency for Research on Cancer (IARC/WHO), "Global cancer observatory (gco): Cancer today," 2024, url: <https://gco.iarc.who.int/en/projects#database>.
- [2] H. Rumgay *et al.*, "Global burden of primary liver cancer in 2020 and predictions to 2040," *J Hepatol*, vol. 77, no. 6, pp. 1598–1606, Dec 2022, doi: [10.1016/j.jhep.2022.08.021](https://doi.org/10.1016/j.jhep.2022.08.021).
- [3] J.-G. Chen *et al.*, "Liver cancer mortality over six decades in an epidemic area: What we have learned," *PeerJ*, vol. 9, p. e10600, 2021, doi: [10.7717/peerj.10600](https://doi.org/10.7717/peerj.10600).
- [4] R. Qiang, X.-Z. Liu, and J.-C. Xu, "The immune pathogenesis of acute-on-chronic liver failure and the danger hypothesis," *Frontiers in Immunology*, vol. 13, 2022, doi: [10.3389/fimmu.2022.935160](https://doi.org/10.3389/fimmu.2022.935160).

- [5] J. Zhou *et al.*, “Guidelines for the diagnosis and treatment of primary liver cancer (2022 edition),” *Liver Cancer*, vol. 12, no. 5, pp. 405–444, Oct 2023, doi: [10.1159/000530495](https://doi.org/10.1159/000530495).
- [6] Y. Liu, A. He, B. Liu, Z. Huang, and H. Mei, “Potential role of lncrna h19 as a cancer biomarker in human cancers detection and diagnosis: A pooled analysis based on 1585 subjects,” *Biomed Res Int*, vol. 2019, p. 9056458, 2019, doi: [10.1155/2019/9056458](https://doi.org/10.1155/2019/9056458).
- [7] X. Chen, J. Song, X. Wang, D. Sun, Y. Liu, and Y. Jiang, “Lncrna linc00460: Function and mechanism in human cancer,” *Thorac Cancer*, vol. 13, no. 1, pp. 3–14, Jan 2022, doi: [10.1111/1759-7714.14238](https://doi.org/10.1111/1759-7714.14238).
- [8] Y. Zhang, “Lncrna-encoded peptides in cancer,” *J Hematol Oncol*, vol. 17, no. 1, p. 66, Aug 2024, doi: [10.1186/s13045-024-01591-0](https://doi.org/10.1186/s13045-024-01591-0).
- [9] M. Yang, H. Lu, J. Liu, S. Wu, P. Kim, and X. Zhou, “Lncrnafunc: A knowledgebase of lncrna function in human cancer,” *Nucleic Acids Res*, vol. 50, no. D1, pp. D1295–D1306, Jan 2022, doi: [10.1093/nar/gkab1035](https://doi.org/10.1093/nar/gkab1035).
- [10] C. Badowski, B. He, and L. X. Garmire, “Blood-derived lncrnas as biomarkers for cancer diagnosis: The good, the bad and the beauty,” *NPJ Precis Oncol*, vol. 6, no. 1, p. 40, Jun 2022, doi: [10.1038/s41698-022-00283-7](https://doi.org/10.1038/s41698-022-00283-7).
- [11] L. Min *et al.*, “Exosomal lncrna rp5-977b1 as a novel minimally invasive biomarker for diagnosis and prognosis in non-small cell lung cancer,” *Int J Clin Oncol*, vol. 27, no. 6, pp. 1013–1024, Jun 2022, doi: [10.1007/s10147-022-02129-5](https://doi.org/10.1007/s10147-022-02129-5).
- [12] Y. Zhang *et al.*, “Lncrna-bc069792 suppresses tumor progression by targeting kcnq4 in breast cancer,” *Mol Cancer*, vol. 22, no. 1, p. 41, Mar 2023, doi: [10.1186/s12943-023-01747-5](https://doi.org/10.1186/s12943-023-01747-5).
- [13] J. Wang, Q. Jia, S. Jiang, W. Lu, and H. Ning, “Pou6f1 promotes ferroptosis by increasing lncrna-casc2 transcription to regulate socs2/slc7a11 signaling in gastric cancer,” *Cell Biol Toxicol*, vol. 40, no. 1, p. 3, Jan 2024, doi: [10.1007/s10565-024-09843-y](https://doi.org/10.1007/s10565-024-09843-y).
- [14] Y. Wu and X. Xu, “Long non-coding rna signature in colorectal cancer: Research progression and clinical application,” *Cancer Cell Int*, vol. 23, no. 1, p. 28, Feb 2023, doi: [10.1186/s12935-023-02867-0](https://doi.org/10.1186/s12935-023-02867-0).
- [15] R. Khalid, H. Naveed, and Z. Khalid, “Computational prediction of disease related lncrnas using machine learning,” *Scientific Reports*, vol. 13, no. 1, p. 806, Jan 2023, doi: [10.1038/s41598-023-27680-7](https://doi.org/10.1038/s41598-023-27680-7).
- [16] M. Li and C. Liang, “Lncdc: A machine learning-based tool for long non-coding rna detection from rna-seq data,” *Scientific Reports*, vol. 12, no. 1, p. 19083, Nov 2022, doi: [10.1038/s41598-022-22082-7](https://doi.org/10.1038/s41598-022-22082-7).
- [17] A. Lawarde, M. Khatun, P. Lingasamy, A. Salumets, and V. Modhukur, “Tumor tissue-of-origin classification using mirna-mrna-lncrna interaction networks and machine learning methods,” *Frontiers in Bioinformatics*, vol. 5, 2025, doi: [10.3389/fbinf.2025.1571476](https://doi.org/10.3389/fbinf.2025.1571476).
- [18] Y. Feng and Y. Wang, “Comparison of the classifiers based on mrna, microrna and lncrna expression and dna methylation profiles for the tumor origin detection,” *Frontiers in Genetics*, vol. 15, 2024, doi: [10.3389/fgene.2024.1383852](https://doi.org/10.3389/fgene.2024.1383852).
- [19] Z. Liu, O. O. Petinrin, M. Toseef, N. Chen, and K.-C. Wong, “Construction of immune infiltration-related lncrna signatures based on machine learning for the prognosis in colon cancer,” *Biochem Genet*, vol. 62, no. 3, pp. 1925–1952, Jun 2024, doi: [10.1007/s10528-023-10516-4](https://doi.org/10.1007/s10528-023-10516-4).
- [20] Y. Liu *et al.*, “Machine learning combined with multi-omics to identify immune-related lncrna signature as biomarkers for predicting breast cancer prognosis,” *Scientific Reports*, vol. 15, no. 1, p. 23863, Jul 2025, doi: [10.1038/s41598-025-10186-9](https://doi.org/10.1038/s41598-025-10186-9).
- [21] J. Wang, J. Xing, and J. Wang, “Predicting mirna-disease association based on hybrid graph autoencoder,” in *2024 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 2024, pp. 4284–4291, doi: [10.1109/BIBM62325.2024.10821965](https://doi.org/10.1109/BIBM62325.2024.10821965).
- [22] Y. Liu, X. Ren, J. Li, X. Chen, and X. Zhu, “Prediction of circrna-drug sensitivity using random auto-encoders and multi-layer heterogeneous graph transformers,” *Applied Intelligence*, vol. 55, no. 4, p. 238, Dec 2024, doi: [10.1007/s10489-024-05859-3](https://doi.org/10.1007/s10489-024-05859-3).
- [23] S. Ning *et al.*, “Lnc2cancer 3.0: An updated resource for experimentally supported cancer-lncrna and circrna associations with web tools,” 2021, url: <https://bio-bigdata.hrbmu.edu.cn/lnc2cancer/>.
- [24] Z. Bao, Z. Yang, Z. Huang, Y. Zhou, Q. Cui, and D. Dong, “Lncrnadisease v3.0: A comprehensive database of long non-coding rna and circular rna associated with human diseases,” 2023, url: <http://www.rnanut.net/lncrnadisease/index.php/home/browse>.
- [25] GENCODE Project, “Gencode human gene annotation, release 47 (grch38.p14),” 2024, url: <https://www.gencodegenes.org/human/>.
- [26] The Cancer Genome Atlas (TCGA) and GDC Data Portal, “Tcga-lihc: The cancer genome atlas liver hepatocellular carcinoma project,” 2024, url: <https://portal.gdc.cancer.gov/projects/TCGA-LIHC>.
- [27] GTEx Consortium, “The genotype-tissue expression (gtex) portal: Adult bulk tissue expression,” 2023, url: https://gtexportal.org/home/downloads/adult-gtex/bulk_tissue_expression.

-
- [28] ENCODE Project Consortium, “Encode encyclopedia of dna elements: Data portal,” 2024, url: <https://www.encodeproject.org/search/>.
 - [29] The UniProt Consortium, “Uniprot id mapping service,” 2024, url: <https://www.uniprot.org/id-mapping>.
 - [30] EMBL-EBI IntAct Team, “Intact: Molecular interaction database,” 2024, url: <https://www.ebi.ac.uk/intact/home>.
 - [31] D. P. Kingma and M. Welling, “Auto-encoding variational bayes,” *arXiv preprint arXiv:1312.6114*, 2013, url: <https://arxiv.org/abs/1312.6114>.
 - [32] W. L. Hamilton, “Graph representation learning,” *Synthesis Lectures on Artificial Intelligence and Machine Learning*, vol. 14, no. 3, pp. 1–159.
 - [33] W. L. Hamilton, R. Ying, and J. Leskovec, “Inductive representation learning on large graphs,” in *Proceedings of the 31st International Conference on Neural Information Processing Systems*, ser. NIPS’17. Red Hook, NY, USA: Curran Associates Inc., 2017, pp. 1025–1035.