Graph Attention Network version two vs Graph Attentiuon Network in Extracting Feature from Simplified Molecular-Input Line-Entry System for HIV Classifition

1st  Agnes Calista   
*Data Science*  
Bina Nusantara *University*Jakarta, Indonesia  
[agnes.calista@binus.ac.id](mailto:agnes.calista@binus.ac.id)

2nd  Muhammad Farrell Hafizh S.   
*Data Science*  
Bina Nusantara *University*Jakarta, Indonesia  
[muhammad.sasongko002@binus.ac.id](mailto:muhammad.sasongko002@binus.ac.id)

3rd  Snow White Putri Safa   
*Data Science*  
Bina Nusantara *University*Jakarta, Indonesia  
[snow.safa@binus.ac.id](mailto:snow.safa@binus.ac.id)

*Abstract*— This research utilizes Simplified Molecular-Input Line Entry System (SMILES) to predict anti-HIV activity. Machine learning algorithms are employed with SMILES representations to differentiate between active and inactive molecules. this research has the same goal as the research by Hugo et.al., that we put the imbalanced data problem aside in this study. However, a challenge lies in handling imbalanced data. To address this issue, we propose using Graph Attention Network version 2 (GATv2) to classify HIV/AIDS instances while disregarding the imbalanced nature of the data. Then split the dataset into three datasets: 70%, 20% and 10% for training set, validation set and test set respectively. We compared our model using GATv2. into However, complete resolution of the imbalanced problem is not achieved in this study. The current solution is to try an oversampling approach for non-tabular datasets. With oversampling, we will duplicate a small amount of data so that the data is balanced and for the upcoming work, we will experiment with this dataset using generative adversarial networks (GANs) technique.

Keywords— HIV, GATv2, SMILES, GNN, HIV Classification

# Introduction

Human Immunodeficiency Virus (HIV) is an infection that attacks the body's cells, weakening the body's immunity to various infections and cancer [1].  HIV targets the body's white blood cells to attack the immune system, making the body vulnerable to diseases such as tuberculosis, infections, and some types of cancer. As the virus damages and destroys the immune system, the infected person becomes immune deficient over time [10].

HIV spreads from the body fluids of an infected person. The fluids are Blood, Semen (cum) and pre-seminal fluid (pre-cum), Rectal fluids, Vaginal fluids, Breast milk. The HIV virus can enter the bloodstream through the mucous membranes in the rectum, vagina, mouth, or tip of the penis when there is direct contact with HIV-infected body fluids, such as semen, vaginal fluid, or anal fluid [13]. In addition to sexual transmission of HIV, use of unsterilized needles is one of the most common routes of transmission especially among injecting drug users. Needle and Syringe Programs (NSP) is a strategy used to prevent HIV transmission in most of the developed countries [14]. HIV is also spread through mothers to their babies. An estimated 1.3 million women and girls living with HIV become pregnant each year worldwide. Without appropriate intervention, the risk of HIV transmission from an HIV-infected mother to her child through pregnancy, childbirth, or breastfeeding is 15-45%. Therefore, it is critical to rapidly identify HIV breakthrough infection and provide lifelong the silent treatment and care, including support for maintenance of congruent care, control of the virus, and provision of partner services[15].

To find out how a person is infected with HIV, there are three types of tests that can be taken. First, antibody test is a rapid test and is the only standalone test approved by the Food and Drug Administration (FDA). It is taken by using blood, oral fluid or blood from a vein. However, the test will be quickly detected by using blood through a vein. Second, antigen/antibody test is used to detect HIV antibodies and antigens. This test is taken in a laboratory and is common in the United States [11]. This laboratory test involves drawing blood from a vein. Third, The Nucleic Acid Test (NAT) is capable of identifying the virus itself in the bloodstream [2], allowing for prompt and precise HIV diagnosis. This procedure entails extracting blood from a vein, sending it to a laboratory, and assessing the presence of HIV or the viral load. It is specifically advised for individuals who have encountered recent or potential exposure, display early symptoms of HIV, and have received negative outcomes from antibody or antigen/antibody tests. To determine the most suitable HIV test for your situation, it is recommended to consult your healthcare provider [11].

The human immune system can never really get rid of HIV. If untreated, HIV will progress to Acquired Immunode ficiency Syndrome (AIDS) [1]. AIDS is the last state of HIV disease that occurs because the immune system has been weakened. When the Cluster of Differentiation 4) CD4 cell count drops below 200 cells per cubic millimeter of blood (200 cells/), it is called AIDS. CD4 cells are a kind of white blood cell. Also called CD4 T lymphocytes or “helper T cells”. That's as long as it helps fight infections by stimulating the immune system to destroy disease-causing viruses, bacteria, and other germs [18]. A normal CD4 count is 500-1,600 cells/) [17].

Although HIV cannot be cured, it can be treated and prevented with antiretroviral therapy (ART) [16].ART uses a combination of HIV drugs or commonly called the HIV therapy program (HIV treatment regimen). the way these medicines work is by reducing the amount of HIV in the body (called viral load). by preventing HIV from reproducing the body's immune system has the opportunity to reproduce more CD4 cells. The effect is that the immune system becomes strong enough to fight breakthrough infection and certain HIV-related cancer. HIV medicines also minimize the risk of HIV transmission [19].

Specifically, this study investigates the implementation of machine learning and deep learning approaches to HIV detection. Machine learning enables analysis of high-dimensional and complex data, enabling accurate pattern recognition and prediction. Furthermore, machine learning algorithms can capture non-linear relationships and interactions between molecular features that are difficult for traditional statistical methods. This ability leads to more accurate predictions and a better understanding of complex molecular phenomena [21].

Which will later be used to develop HIV drugs. Usually they use protein-based datasets. But in this research using Simplified Molecular-Input Line Entry System (SMILES). SMILES strings are the most concise text-based molecular representations. The SMILES method is very useful in drug development because it allows to visually represent the molecular structure and predict its chemical properties. SMILES was used as part of the dataset to distinguish between molecules with potential anti-HIV activity and those without. Machine learning algorithms were used in conjunction with SMILES to predict the anti-HIV activity of these molecules. [20].

The objective of this research is to identify cases of HIV/AIDS by utilizing the Simplified Molecular-Input Line Entry System (SMILES) as input data. Then this research has the same goal as the research by Hugo et.al., that we put the imbalanced data problem aside in this study. The proposed approach involves employing Graph Attention Network version 2 (GATv2) as the classification model for distinguishing instances of HIV/AIDS. Then to prove that GATv2 can handle the imbalanced problem by ignoring the fact that our data is imbalanced [6].

# Related work

There are several studies that are relevant to this research topic. In related research by Gasteiger and friends (2021) it is said that graph neural networks (GNNs) has demonstrated considerable promise in their ability to predict molecular energy and various quantum mechanical properties. These networks exhibit faster prediction capabilities in comparison to traditional quantum chemistry approaches, all while maintaining a similar level of accuracy. Consequently, GNNs facilitate the simulation of larger systems with enhanced precision. Nevertheless, GNNs still encounter significant theoretical and practical constraints that need to be addressed. GNNs can be used for data molecules because it can train models to study patterns and interactions in complex molecular structures, as well as utilize information from the immediate surroundings to predict molecular properties or behaviors, such as biological activity, stability, or chemical properties [8].

According to a research written by Brody, Alon, and Yahav in (2021), it is stated that GATv2 surpasses extensively tuned GNNs, achieving a significant performance boost of more than 1.4% on the challenging "UnseenProj Test" set of the VarMisuse task without requiring hyperparameter tuning. Other than that, GATv2 demonstrates an impressive 11.5% improvement over an extensively tuned GAT across 13 prediction objectives in the QM9 dataset [3].

# Methodology

This research has the same goal as the research by Hugo et.al., that we put the imbalanced data problem aside in this study. We intentionally did not use the sampling method because we wanted to prove that graph attention network (GAN) can provide better analysis and classification on imbalanced data. Therefore, we did not apply the data balancing process in this study. However, our main difference lies in the choice of model used. While the research by Hugo et.al. adopts the graph convolutional network (GCN) and GAT models, we choose to use the GNN and the latest version of the Graph Attention Network (GATv2). We use GATv2 because GATs use a static attention mechanism, there are simple graph problems that GAT cannot express: in a controlled problem, we show that static attention hinders GAT from even fitting the training data. To remove this limitation, we decided to use GATv2 [3].

## Dataset Description

The data was collected by National Cancer Institute from the Drug Therapeutics Program (DTP) AIDS Antiveral Screen. The dataset consists of 41,1127 instances with active and non-active categories.

Sebuah gambar berisi teks, cuplikan layar, Plot, garis

Description automatically generated

Fig. 1. Data Distribution

According to figure 1, active data has a frequency of 39,684 meanwhile non-active data has a frequency of 1,442. This distribution of the HIV dataset indicates an imbalance. However, for this experiment, we deliberately chose to overlook the issue of imbalance. The reason we ignored the data balancing process is because we want to prove that the Graph Attention Network version 2 would provide better analysis and classification on the imbalanced data.

## Proposed Methodology

In this paper, we want to know the performance of GNN-GATv2. GNN-GATv2, is based on a multi-layer architecture. The core components include GATv2Conv layers, TopKPooling operations, and Linear transformations.

To do that we compared the performance of GAT and GATv2 with HIV dataset, in which the dataset have an imbalance problem. We use Gatv2 because GATs use a static attention mechanism, there are simple graph problems that GAT cannot express: in a controlled problem, we show that static attention hinders GAT from even fitting the training data. To remove this limitation, we decided to use GATv2 [3].

We split the dataset into three datasets: 70%, 20% and 10% for training set, validation set and test set respectively. We compared our model using GATv2.

We then employ three GATv2Conv layers with dropout regularization to capture local and global graph information. TopKPooling is utilized to select informative nodes based on their attention scores. The resulting node representations are concatenated and passed through linear layers for classification.

## Graph Neural Networks (GNNs)

In this study, Graph Neural Networks (GNN) were used to process molecule data as it can train models to study patterns and interactions in complex molecular structures and utilize information from the surrounding environment predict molecular properties or behaviors, such as biological activity, stability, or chemical properties [8]. GNN updates each layer of each nodes’ representation by combining the representations of its neighboring nodes.

Sebuah gambar berisi Font, tulisan tangan, putih, teks

Description automatically generated(1)

is the value of the update node, and Sebuah gambar berisi Font, tipografi, kaligrafi, deasin

Description automatically generated is the messages coming to node calculate ahead. Then using the same 2 neural networks on and for each vertex that made up the graph [22].

Modern GNNs are based on an aggregating messaging layer. It iteratively updates each node with neighbor information. This framework was first proposed by Gilmer et al [23].

Sebuah gambar berisi Font, teks, putih, tulisan tangan

Description automatically generated(2)

Formula (2), Sebuah gambar berisi Font, sketsa, papan klip, Grafis

Description automatically generated Computing and Aggregating Messages from Adjacent Nodes and Edges.

Sebuah gambar berisi Font, putih, teks, tipografi

Description automatically generated (3)

Formula (3), Sebuah gambar berisi Font, sketsa, papan klip, Grafis

Description automatically generated are update node functions based on messages and old functions.

Repeating the two operations updates the node properties on the diagram and makes them available for node-level tasks. B. Molecular property prediction for performing graph-level tasks such as: Different graphs have different numbers of knot features, so it may not be feasible to use all knot features directly.

Sebuah gambar berisi Font, putih, Grafis, papan klip

Description automatically generated express the feature vector, Sebuah gambar berisi Font, putih, kaligrafi, tipografi

Description automatically generated. Of nodeSebuah gambar berisi Font, tipografi, kaligrafi

Description automatically generated the k-th level. Models all neighbors of node Sebuah gambar berisi Font, tipografi, kaligrafi

Description automatically generated. Sebuah gambar berisi Font, tipografi, kaligrafi

Description automatically generated and Sebuah gambar berisi kaligrafi, tipografi

Description automatically generated is Sebuah gambar berisi Font, kaligrafi, putih, tipografi

Description automatically generated

Each node Sebuah gambar berisi Font, tipografi, kaligrafi

Description automatically generated is initialized as feature vector Sebuah gambar berisi Font, tipografi, teks, putih

Description automatically generated

through the node insertion function.

Each node is assigned a feature vector via an edge embedding function Sebuah gambar berisi Font, teks, kaligrafi, tipografi

Description automatically generated [24].

## Graph Attention Network version 2 (GATv2)

Graph Attention Networks (GATs) are one of the most popular Graph Neural Network (GNN) architectures and are considered as the state-of-the-art architecture for representation learning with graphs [3]. GAT have a restricted kind of attention as static attention and distinguish it from a strictly more expressive dynamic attention. Because GATs use a static attention mechanism, there are simple graph problems that GAT cannot express: in a controlled problem, we show that static attention hinders GAT from even fitting the training data. To remove this limitation, we decided to use GATv2 [3].

GATv2 is a dynamic graph attention variant that is strictly more expressive than GAT [3]. GATv2 work on graph based data similar to GAT. A graph is composed of vertices and edges connected to the vertices. The GATv2 operator solves the static attention problem of the standard GAT. The static attention refers to when the attention on key vertices has the same rank (order) for each query vertex [8].

A single layer of graph attention v2 is present in the model. GATv2 consists of several layers of the same type. It receives as input a set of node representations and outputs transformed representations. }

as input and outputs }, which is

Operations within an attention mechanism in a GATv2. Specifically, they can be described as follows:

1. Compute the number of proportion through head. This step typically refers to the number of attention heads used in the graph neural network. Each head performs its own computation independently, allowing the network to capture different aspects of the input data.
2. Linear layer of initial source transformation. This step, transform the source node's embeddings before self-recognition.
3. Linear layer for computing attention values ​​ In this step, another linear layer is used to compute the attention values between nodes in the graph. These values represent the importance or relevance of each target node j to a given source node i.
4. Activation of Attention Score . The attention scores obtained from the previous step are typically passed through an activation function to introduce non-linearity and capture complex relationships in the graph. Common activation functions include sigmoid, tanh, or ReLU.
5. Softmax for calculating attention Softmax normalizes the scores, ensuring they sum up to 1 and represent a valid probability distribution over the target nodes. This allows the network to assign relative importance to each target node.
6. Dropout layer applicable for attention.
7. The first transformation,

(4)

perform two linear transformations on each head and split them head by head.

1. Next, compute the attention value for each head k. neglect the ⋅k for simplicity.

, (5)

is attention value from node to node . Compute this for each head.

is a mechanism of attention, which computes the attention value.

1. Then calculate for all the pairs of.
2. Normalize attention coefficient

(6)

In which is of the set of I-connected nodes. Perform these by setting the unlinked to - that makes the exp () ∼ 0 for non-connected pair.

1. Implementing dropout regulation within neural network while considering the given steps and equations related to attention mechanisms.
2. calculate the final output for each head in your neural network's forward pass, considering the steps and equations mentioned earlier in the context of the attention mechanism.

= (7)

1. Define the concatenation of the head outputs in your neural network's forward method.

(8)

1. To take the mean of the outputs from each head Define the mean calculation of the head outputs in your neural network's forward method:

(9)

# Result and Discussion

## Experiment Setting

For creating GAT, we utilized PyTorch Geometry [5] library. All the experiments were conducted on Google Colab, and we employed GPU acceleration for training.

PyTorch is a very popular framework in the field of machine learning and artificial intelligence. It has several key advantages that set it apart from other frameworks. PyTorch uses a dynamic computational model that allows developers to build and change models interactively, making debugging and experimentation easier. The PyTorch user community is also very active, providing access to a wide range of learning resources and extensive support [5].

## Result GATv2

Table I through Table III display the outcomes of GATv2 when using different weights. Table I, II, and III contain the results of the training, validation, and testing process.

1. GATv2 training evaluation metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Weight** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| (1, 1) | 3.70% | 0.49931 | 0.48519 | 0.035911 |
| (0.5 , 1) | 4.01% | 0.50092 | 0.50674 | 0.039199 |
| (-10, 1) | 94.42% | 0.49937 | 0.49899 | 0.49835 |
| (-1, 1) | 96.36% | 0.50007 | 0.50302 | 0.49168 |

1. gatv2 validation evaluation metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Weight** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| (1, 1) | 3.45% | 0.5 | 0.017264 | 0.033376 |
| (0.5 , 1) | 3.45% | 0.5 | 0.017264 | 0.033376 |
| (-10, 1) | 96.30% | 0.49868 | 0.48269 | 0.49055 |
| (-1, 1) | 96.55% | 0.5 | 0.48274 | 0.49122 |

1. gatv2 testing evaluation metrics

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Weight** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| (1, 1) | 3.23% | 0.5 | 0.016168 | 0.031324 |
| (0.5 , 1) | 3.23% | 0.5 | 0.016168 | 0.031324 |
| (-10, 1) | 96.40% | 0.49812 | 0.48377 | 0.49084 |
| (-1, 1) | **96.77%** | 0.5 | 0.48383 | 0.49178 |

The occurrence of a mismatch between precision and recall, as observed in Table II and Table III, happened due to class imbalance. This happen because we ignored the imbalance problem in the dataset. We choose to ignore the imbalance problem because we want investigate if GATv2 can handle imbalance dataset problem.

Veličković the author of GAT, has confirmed on Twitter that GAT was designed to work in the “easy-to-overfit” datasets of the time (2017), such as Cora, Citeseer and Pubmed (Sen et al., 2008), where the data might had an underlying static ranking of “globally important” nodes. Veličković agreed that newer and more challenging benchmarks may demand stronger attention mechanisms such as GATv2. That is why we choose to propose GATv2 in this study to investigate does GATv2 can perform better than GAT in handling imbalanced dataset.

1. Comparison GATv2 and GAT

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Method** | **Accuracy** | **Precision** | **Recall** | **F1-score** |
| GATv2 | 96.77% | 0.5 | 0.48383 | 0.49178 |
| GAT | 96.50% | 0.5 | 0.48249 | 0.49109 |

From table IV we see that both model have a slight difference in F1-score. F1-score is the harmonic mean of precision and recall and is a better measure than accuracy. It can be summarized that GATv2 performed slightly better than GAT in terms of handling imbalance dataset.

# Conclusion

In this paper we want to know the performance of GATv2, to accomplish that we tried to classifiy HIV/AIDS by using Simplified Molecular-Input Line-Entry System (SMILES) using the Graph Attention Network version 2 (GATv2) and Graph Attention Network model in which the data was collected by National Cancer Institute from the Drug Therapeutics Program. After we tried to construct the model, we then compare the two model using confusion matrix which include Accuracy, Precision, Recall, and F1-score. Based on the result of our experiment, GATv2 gave slightly better performance in handling imbalance dataset. But GATv2 cannot handle imbalanced data problem and it can be seen through the result of our experiment in which the result is biased. To conclude all of the statements that have been stated, GATv2 does perform better than GAT in handling imbalanced data as seen form Table IV that GATv2 have better accuracy which is 96,7% than GAT which have 96,50% accuracy. Though GATv2 does not perform well on imbalanced data, and further research is needed.

##### References

1. World Health Organization. (2019). HIV/AIDS. Fact Sheet. <https://www.who.int/news-room/fact-sheets/detail/hiv-aids> (accessed June. 17, 2023)
2. UCSF Health. (n.d.). HIV Diagnosis. Accessed from <https://www.ucsfhealth.org/conditions/hiv/diagnosis>
3. Brody, S., Alon, U., & Yahav, E. (2021). How attentive are graph attention networks?. *arXiv preprint arXiv:2105.14491*.
4. LabML. (n.d.). Graph Attention Network v2 (GATv2). Accessed on [Mei, 28 2023], from <https://nn.labml.ai/graphs/gatv2/>
5. Mohammed, R., Rawashdeh, J., & Abdullah, M. (2020, April). Machine learning with oversampling and undersampling techniques: overview study and experimental results. In *2020 11th international conference on information and communication systems (ICICS)* (pp. 243-248). IEEE.
6. M. Fey and J. E. Lenssen, “Fast Graph Representation Learning with PyTorch Geometric,” 2019, doi: 10.48550/ARXIV.1903.02428.
7. Khan, M.A.I. (2023, May 3). Activation Functions in PyTorch. Deep Learning with PyTorch. Accessed on (June 10, 2023) from <https://machinelearningmastery.com/activation-functions-in-pytorch/>
8. Gasteiger, J., Becker, F., & Günnemann, S. (2021). Gemnet: Universal directional graph neural networks for molecules. Advances in Neural Information Processing Systems, 34, 6790-6802.
9. <https://arxiv.org/pdf/2105.14491.pdf>
10. World Health Organization. (2023). HIV/AIDS. <https://www.who.int/health-topics/hiv-aids/#tab=tab_1> (accessed June. 17, 2023)
11. Centers for Disease Control and Prevention. (2022). HIV Testing Types. Retrieved June 17, 2023, from <https://www.cdc.gov/hiv/basics/hiv-testing/test-types.html>
12. Morris, M., & Kretzschmar, M. (1997). Concurrent partnerships and the spread of HIV. Aids, 11(5), 641-648.
13. HIV.gov. (2022, June 16). How is HIV Transmitted? Retrieved from <https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/how-is-hiv-transmitted/>
14. MacDonald, M., Law, M., Kaldor, J., Hales, J., & Dore, G. J. (2003). Effectiveness of needle and syringe programmes for preventing HIV transmission. International Journal of Drug Policy, 14(5-6), 353-357.
15. World Health Organization. (2023). Mother-to-child transmission of HIV.https://www.who.int/teams/global-hiv-hepatitis-and-stis-programmes/hiv/prevention/mother-to-child-transmission-of-hiv (accessed June. 17, 2023)
16. Patel, P., Rose, C. E., Collins, P. Y., Nuche-Berenguer, B., Sahasrabuddhe, V. V., Peprah, E., ... & Operating, N. P. D. C. T. (2018). Noncommunicable diseases among HIV-infected persons in low-income and middle-income countries: a systematic review and meta-analysis. AIDS (London, England), 32(Suppl 1), S5.
17. HIV.gov. (2023, january 13) What Are HIV and AIDS? Retrieved June 17, 2023 from <https://www.hiv.gov/hiv-basics/overview/about-hiv-and-aids/what-are-hiv-and-aids/>
18. MedlinePlus. (2023, August 22). CD4 Lymphocyte Count. Accessed on [June, 17 2023], from <https://medlineplus.gov/lab-tests/cd4-lymphocyte-count/>
19. HIV.gov. (2023, August 16). HIV Treatment. Accessed on [June, 18 2023], from https://hivinfo.nih.gov/understanding-hiv/fact-sheets/hiv-treatment-basics
20. Hugo, G., Sinaga, V. L., Dinata, I. M., Kurniadi, F. I., & Anggreainy, M. S. (2022, September). Graph Attention Network on Extracting Feature from Simplified Molecular-Input Line-Entry System for HIV Classification. In 2022 International Conference on Electrical and Information Technology (IEIT) (pp. 398-401). IEEE.
21. Z. Wu et al., “MoleculeNet: a benchmark for molecular machine learning,” Chem. Sci., vol. 9, no. 2, pp. 513–530, 2018, doi: 10.1039/C7SC02664A
22. Dagli, R. (February 1, 2022). "What Are Graph Neural Networks? How GNNs Work, Explained with Examples." Retrieved <https://www.freecodecamp.org/news/graph-neural-networks-explained-with-examples/>
23. Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O. & Dahl, G. E. Neural message passing for quantum chemistry in International conference on machine learning (2017), 1263–1272.
24. Wang, Y., Wang, J., Cao, Z., & Barati Farimani, A. (2022). Molecular contrastive learning of representations via graph neural networks. Nature Machine Intelligence, 4(3), 279-287.
25. Brody, S., Alon, U., & Yahav, E. (2021). How attentive are graph attention networks?. arXiv preprint arXiv:2105.14491.