HIV Classification from Simplified Molecular-Input Line-Entry System using Random Forest Algorithm

Christopher Stanley
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
christopher.stanley001@binus.ac.id

Joselyn Setiawan
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
joselyn.setiawan@binus.ac.id

Jackson Bhakti Gusman
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
jackson.gusman@binus.ac.id

Felix Indra Kurniadi
Computer Science Department, School of Computer Science
Bina Nusantara University
Jakarta, Indonesia
felix.indra@binus.ac.id

***Abstract*—Human immunodeficiency virus (HIV) targets the immune system as its main target, which impairs the body's ability to fight off infections. Protein-based methods have been the main focus of current HIV detection research. However, comprehending protein structures is difficult due to their intricacy. The Simplified Molecular-Input Line-Entry System (SMILES) is used as our input in this paper as we suggest a distinct strategy. Comparing SMILES to protein structures, the analysis is made simpler by the linear structure it offers. In our study, the Random Forest technique is used to extract features from SMILES representations. We contrast our approach with RDKIT, a well-known SMILES feature extraction library, to see how effective the result will be. According to the results, the suggested Random Forest algorithm obtains a remarkable accuracy rate of 99.82%. However, it is worth nothing that this approach has limitations when handling imbalanced data.**

Keywords—HIV, SMILES, Random Forest, Machine Learning, RDKIT

I. Introduction

HIV is a virus that specifically attacks the immune system, making it less effective in fighting off various infections and cancers that a healthy immune system can typically handle. As the virus progressively damages and hinders the functionality of immune cells, individuals infected with HIV experience a gradual decline in their immune defenses. Without appropriate treatment, it may take several years for the infection to progress to the most advanced stage, known as acquired immunodeficiency syndrome (AIDS) [1].

The symptoms of HIV differ depending on the stage of infection. While individuals are generally highly contagious during the initial months after contracting HIV, many remain unaware of their status until later stages. In the first few weeks after the initial infection, some individuals may not experience any symptoms or may have flu-like symptoms, including fever, headache, rash, and a sore throat [1]. Researchers have made significant efforts to detect HIV/AIDS using various approaches, including machine learning, data mining, and deep learning. Previous studies have predominantly focused on protein-based datasets [2]-[4].However, there has been limited exploration of using the Simplified Molecular-Input Line Entry System (SMILES) as a dataset for HIV/AIDS detection. While proteins may provide more comprehensive information about HIV/AIDS, it is worth noting that molecular information can offer enhanced analysis for detecting HIV/AIDS. Therefore, utilizing SMILES as a dataset presents an opportunity to gain valuable insights and improve the accuracy of HIV/AIDS detection methods.

In this paper, we classified HIV using SMILES. We proposed a random forest algorithm for extracting SMILES features. The proposed method is compared with the feature extractor from RDKIT [5]–[7].

II. Literature Review

HIV classification is a machine learning task that involves predicting the type of HIV strain from a given set of features. Many studies have explored many strategies and algorithms to find the best method for HIV classification.

Abdullah et al. (2021) used SMILES strings to represent the chemical structures of HIV strains and achieved 97% accuracy in classifying them using the Random Forest algorithm [2]. Lu et al. (2018) explored deep learning techniques to predict HIV-1 protease cleavage sites, revealing their potential for understanding molecular interactions related to HIV infection [3]. Hu et al. (2020) emphasized considering dynamic relationships between molecular components to enhance the prediction of HIV-1 protease cleavage sites [4].

Achary et al. (2019) investigated using graph invariants and SMILES attributes to model molecular properties, demonstrating the applicability of SMILES-based approaches [5]. Raposo et al. (2020) employed the Random Forest algorithm to predict HIV drug resistance mutations, aiding treatment decisions [6]. Zhang et al. (2022)

highlighted the ability of Random Forest algorithms to handle multiclass imbalance issues, relevant for accurate HIV classification [8]. Li et al. (2018) applied machine learning methods to classify HIV-1 protease inhibitors, contributing to drug discovery efforts [9]. Morris et al. (2018) showcased the utility of convolutional neural networks in understanding interactions between molecular structures and HIV-1 proteins [10].

These studies collectively demonstrate the potential of machine learning, deep learning, and random forest algorithms in HIV classification. By analyzing molecular data and protein interactions, these approaches can improve accuracy and efficiency in detecting HIV, aiding disease management and prevention strategies.

III. Methodology

A. Dataset

The HIV/AIDS dataset that is used in this paper was collected from the Drug Therapeutics Program (DTP) AIDS Antiveral Screen. The dataset consists of 41,127 rows and 3 columns or features. The features are then grouped into 2 classes. The column Confirm Active (CA) and Confirmed Moderately Active (CM) belongs to the 'active' class. The column Confirmed Inactive (CI) belongs to the 'non-active' class. Figure 1 shows the distribution of the data.

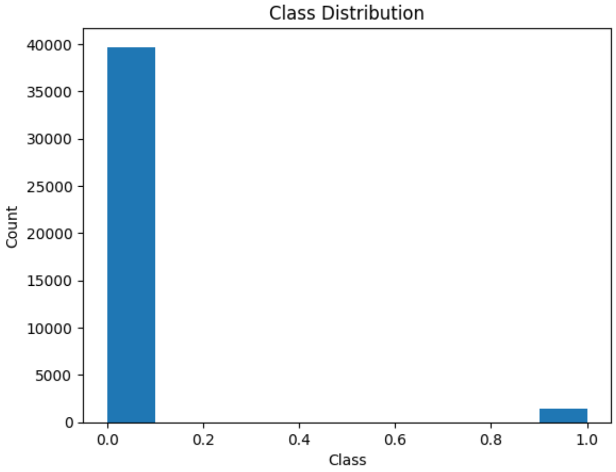

Fig. 1. The Distribution of the Data

The histogram in figure 1 shows that the distribution of the data is not uniform between both classes. This large difference between classes could lead to misleading results. Therefore in this paper, we proposed a solution to handle the class imbalance using a technique called resampling.

B. Resampling

Resampling refers to a technique which tries to create more balance between classes [7]. There are two types of resampling. The first one is called oversampling. Oversampling could be done by creating more observations from minority class. The second is called undersampling. Undersampling could be done by only including a subset sample from the majority class. In this paper, we carried out an oversampling technique using RandomOverSampler imported from the sklearn library.

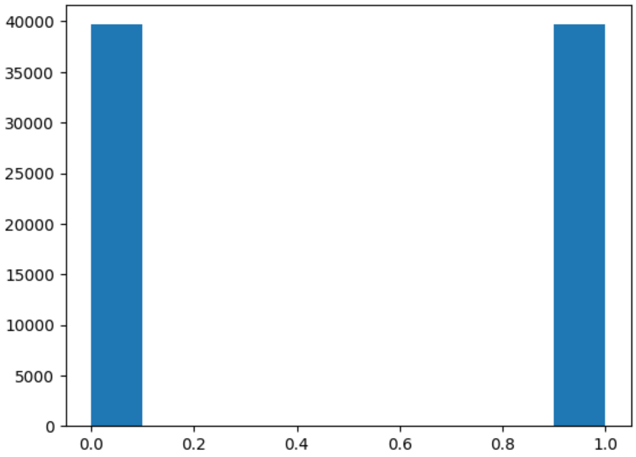

Fig. 2. The Distribution of the Data after Resampling

After applying oversampling as in figure 2, we can see that both classes now have equal numbers of data. This helps increase the performance of the classifier later on the model.

C. Feature Extraction

Extracting features from graph data can be challenging as it is different from tabular data or any other data. With each node being connected, we can't miss the information it contains. For the feature extraction step, we use RDKit, NetworkX, and Graph2Vec.

RDKit is an open-source cheminformatics library that is commonly used to deal with chemical data and structure. In this project, we use RDKit to convert SMILES into RDKit molecules. The aim of this conversion is to enable us to use several functions that help us to interact with these molecules.

Graph2Vec is an algorithm or technique used for graph embedding, which involves transforming a graph into a vector representation. In the provided context, the RDKit molecules are converted into a graph representation using the function mol_to_nx, resulting in a NetworkX graph object. This graph representation can then be utilized for various graph processing tasks or fed into a graph-based model for further analysis or prediction. Graph2Vec captures the structural and topological information of the graph, allowing for meaningful vector representations that can be used in downstream graph processing tasks or machine learning applications.

D. Random Forest

Random forest is an ensemble learning method that constructs a multitude of decision trees at training time and outputs the class that is the mode of the classes output by individual trees. Random forest is a powerful machine learning algorithm that can be used for a variety of tasks, including classification, regression, and outlier detection. In this research we used a random forest algorithm to train a classifier on a dataset of SMILES strings and HIV protein sequences. Random forest is quite well-known for its performance because the tree-based model is known to handle class imbalance better than any other model [8]. The architecture of random forest is shown in figure 4.

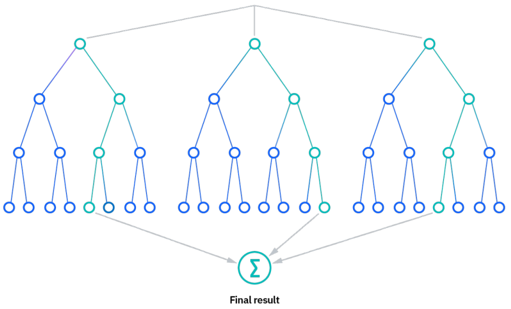

Fig. 4. Architecture of a Random Forest

IV. Experiment and Result

Resampling is a technique that can be used to reduce overfitting in random forest. Resampling involves creating multiple training and test sets from the original data. This allows the model to be trained and evaluated on multiple datasets, which can help to improve its generalization performance. Therefore, if the dataset is imbalanced, we have to use a resampling technique before training the random forest model. By resampling the data to balance the classes, we can ensure that the model doesn't create bias towards the majority class. This helps to address the issue of class imbalance and improve the model's ability to generalize and make accurate predictions for both classes.

In this experiment, we will compare the evaluation metrics before and after applying the oversampling resampling technique to three different models: Random Forest, SVM, and Logistic Regression. The aim is to assess the impact of the resampling technique on the performance of these models when dealing with an imbalanced dataset.

By measuring the evaluation metrics, such as accuracy, precision, recall, and F1-score, we can evaluate and compare the performance of each model before and after resampling. These metrics provide insights into the models' ability to correctly classify instances from both the majority and minority classes.

The following steps were taken to conduct the experiment:

1. Split the dataset into a training and test set with a 80/20 split.
2. Each models were trained on the training set using original imbalanced data.
3. Calculate the evaluation metrics on models before applying oversampling.
4. Applied Oversampling technique to minority class
5. Each models were trained on the training set using resampled data.
6. Calculated the evaluation metrics on models after applying oversampling.

Results are shown in the table below:

Table I. Comparison Result of Different Method

Model	Accuracy	Precision	Recall	F1-Score
Random Forest	96%	80%	10%	18%
Random Forest (Resampled)	99%	99%	100%	99%
Logistic Regression	96%	80%	5%	2%
Logistic Regression (Resampled)	64%	68%	54%	60%
SVM	96%	91%	7%	13%
SVM (Resampled)	75%	84%	60%	70%

The first model listed is the Random Forest model. It achieves an accuracy of 96% and a precision of 80%, but it exhibits a low recall of 10% and an F1-score of 18%, indicating a bias towards false negatives. This means the model struggles to identify positive instances correctly. After applying oversampling, the Random Forest model with resampling demonstrates significant improvements in all metrics. It achieves an accuracy of 99% and precision of 99%, accompanied by a perfect recall of 100% and an impressive F1-score of 99%. This suggests that the model is now capable of accurately capturing positive instances, addressing the previous bias observed in the initial model. The use of resampling techniques has effectively enhanced the model's performance, resulting in a more balanced and accurate classification outcome.

Next model listed is the Logistic Regression model. This model has an accuracy of 96% and a precision of 80%. However, it has low recall of 5% and an F1-score of only 2%, which is the same case as Random Forest Model. But, even after applying oversampling, the Logistic Regression Model still has some limitations. It has a lower accuracy of 64% and a precision of 68%, with a modest recall of 54% and an F1-score of 60%. These results show that the oversampling helps to alleviate the bias to some extent, but it is not enough to fully correct the model's inability to accurately identify positive instances.

Support Vector Machine is the last model on the list. The accuracy and precision of the model are 96% and 91%, respectively. However, it only had a 7% recall and a 15% F1-score. This model produces bias toward false negatives in the same manner as Random Forest and Logistic Regression models. The Support Vector Machine has some limitations when applying oversampling, including a reduced accuracy of 75% and a precision of 84%, as well as a modest recall of 60% and an F1-score of 70%. Similar to

Logistic Regression, the results demonstrate that oversampling partially corrects the bias but does not entirely eliminate the model's failure to correctly detect positive examples.

Overall, the experiment showed that resampling can improve the performance of machine learning models on imbalanced datasets, but the improvement is model-dependent. The Random Forest model showed the greatest improvement after resampling, while the Logistic Regression model showed the least improvement. This suggests that the choice of model may be important when dealing with imbalanced datasets.

Further research is needed to develop more effective methods for dealing with imbalanced datasets. Resampling is a promising approach, but it is not the only approach. Other approaches, such as data augmentation and ensemble learning, may also be effective.

A. Random Forest Result

Table II. Random Forest Evaluation Metrics

Accuracy	Precision	Recall	F1 Score
99.8%	99.8%	99.8%	99.8%

Table III. Confusion Matrix

	Actual Values	
Predicted Values	7902	29
	0	7943

Table 2 and 3 showed outstanding performance by random forest. It is indicated by all the evaluation metrics values are close to 100%. High evaluation metrics means the better the model is at predicting 0 classes as 0, 1 classes as 1, and good at discriminating between individuals with HIV and those who do not have HIV.

The confusion matrix shows that the model correctly classified 7902 out of 7931 HIV-positive SMILES strings and 7943 out of 7943 HIV-negative SMILES strings. This indicates that the model is very good at distinguishing between HIV-positive and HIV-negative SMILES strings.

V. Conclusion

In this paper, an ensemble learning method, Random Forest, is evaluated for HIV classification using Simplified Molecular-Input Line-Entry System (SMILES). The proposed method was evaluated on an imbalanced class dataset of 41127 SMILES strings. The proposed method performed well after carrying out a resampling technique using oversampling the minority class. The results showed that it achieved an accuracy of 99.81% and an F1-score of 99.82% which makes the proposed method superior to logistic regression and support vector classifier. It is also proven that the proposed method was able to correctly classify all of the HIV-positive samples in the dataset.

The proposed method could be used to improve the diagnosis of HIV infection. It could also be used to monitor the effectiveness of HIV treatment. Additionally, it could be used to identify new sources of HIV infection.

The proposed method is still under development, but the results of this study suggest that it has the potential to be a valuable tool for HIV research and prevention.

In the future, we plan to improve the performance of the proposed method by using a larger dataset and by incorporating additional features into the model. We also plan to evaluate the proposed method on a clinical dataset to assess its potential for use in the diagnosis of HIV infection.

References

[1] WHO, "HIV/AIDS," Nov. 30, 2021. https://www.who.int/news-room/fact-sheets/detail/hiv-aids (accessed May. 30, 2023).

[2] M. Abdullah, S. A. Khawaja, and M. Farooq, "HIV-1 Protease Cleavages," in 2021 International Conference on Innovative Computing (ICIC), Lahore, Pakistan, Nov. 2021, pp. 1–7. doi: 10.1109/ICIC53490.2021.9692978.

[3] X. Lu, L. Wang, and Z. Jiang, "The Application of Deep Learning in the Prediction of HIV-1 Protease Cleavage Site," in 2018 5th International Conference on Systems and Informatics (ICSAI), Nanjing, Nov. 2018, pp. 1299–1304. doi: 10.1109/ICSAI.2018.8599496.

[4] L. Hu, P. Hu, X. Luo, X. Yuan, and Z.-H. You, "Incorporating the Coevolving Information of Substrates in Predicting HIV-1 Protease Cleavage Sites," IEEE/ACM Trans. Comput. Biol. Bioinform., vol. 17, no. 6, pp. 2017–2028, Nov. 2020, doi: 10.1109/TCBB.2019.2914208

[5] P. G. R. Achary, A. P. Toropova, and A. A. Toropov, "Combinations of graph invariants and attributes of simplified molecular input-line entry system (SMILES) to build up models for sweetness," *Food Research International*, vol. 122, pp. 40–46, Aug. 2019, doi: 10.1016/j.foodres.2019.03.067.

[6] L. M. Raposo, P. F. F. Rosa, and F. F. Nobre, "Random Forest Algorithm for Prediction of HIV Drug Resistance," in *STEAM-H: Science, Technology, Engineering, Agriculture, Mathematics & Health*, Springer International Publishing, 2020, pp. 109–127. doi: 10.1007/978-3-030-38021-2_6.

[7] U. Pujianto, M. I. Akbar, N. T. Lassela, and D. Sutaji, "The effect of resampling on Classifier Performance: An empirical study," *Knowledge Engineering and Data Science*, vol. 5, no. 1, p. 87, 2022. doi:10.17977/um018v5i12022p87-100

[8] Zhang, L., Sun, P., Huettmann, F., & Liu, S. (2022). The competence of random forests (RF) to deal with multiclass imbalance issue. In Proceedings of the 2022 ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), New York, NY, USA, August 12–18, 2022, pp. 3086–3095. Association for Computing Machinery, New York, NY, USA, 2022. ISBN 978-1-4503-9360-6. doi: 10.1145/3494884.3494943

[9] Y. Li, Y.-J. Tian, Z. Qin, and A. Yan, "Classification of HIV-1 Protease Inhibitors by Machine Learning Methods," *ACS Omega*, vol. 3, no. 11, pp. 15837–15849, Nov. 2018, doi: 10.1021/acsomega.8b01843.

[10] P. Morris, Y. DaSilva, E. Clark, W. C. Hahn, and E. Barenholtz, *Convolutional Neural Networks for Predicting Molecular Binding Affinity to HIV-1 Proteins*. 2018. doi: 10.1145/3233547.3233596