Enhancing HIV Protease Cleavage Site Identification: A Comparative Analysis of F1 Score, NPV, and MCC

Navneet Kaur
Research Scholar, AGC, Amritsar
navkaur9955@gmail.com

Abstract: HIV and its most severe manifestation, AIDS, continue to be a major global health concern. Understanding the proteolytic activities of HIV's protease enzyme is essential for developing effective strategies to combat viral progression and transmission. Although researchers have previously created antiretroviral therapies and inhibitors, issues with toxicity and limited availability persist. This paper examines the current state of predictive models designed to identify protease cleavage sites in HIV-I AIDS proteins, offering an overview of the employed methodologies, data, and existing challenges. By reviewing published works and methodologies to date, this paper aims to provide insights into the present capabilities of machine learning models, specifically DCNN, and potential future advancements in predicting protease cleavage sites for HIV-I AIDS. Additionally, we propose a novel approach that integrates feature extraction and classification using machine learning techniques. The research objective is to conduct a comprehensive analysis of confusion matrix performance metrics, including NPV, F1 Score, and MCC, which are utilized to evaluate machine learning model performance in binary classification tasks.

Keywords: F1 Score, MCC, Negative, Predictive Value.

1. Introduction

AIDS relies on HIV-1 protease, a vital enzyme for replication. This enzyme functions at specific active sites on its surface. Small molecules, known as HIV-1 protease inhibitor drugs, bind to this site, disrupting the normal operation of the enzyme (Gulnik et al., 2000). Understanding and predicting HIV-1 protease cleavage sites in proteins is crucial because cleaved substrates serve as models for developing tightly bound, chemically modified inhibitors. This cleavage process represents a significant irreversible post-translational modification that plays a key role in numerous physiological processes. Many diseases stem from a protease imbalance. Notably, several proteases exhibit specificity, cleaving only the target solvents with particular structural compositions and amino acid residue sequence patterns. Thus, understanding the substrate cleavage specificity for individual proteases is essential for understanding protease functional activity. Substrate specificity can be assessed using peptide specificity profiling or mass spectrometry-based high-throughput methods, each of which has its own strengths and weaknesses.

Given that the experimental identification of protease cleavage events is challenging, costly, and time intensive, the development of effective computational methods and tools to complement experimental approaches is highly beneficial. AIDS poses a significant threat to sustainable development, primarily owing to its global spread and lack of curative treatment. AIDS therapy employs three main approaches: integrase, HIV protease, and reverse transcriptase inhibitors (Ghosh et al., 2016).

The pharmaceutical industry focuses primarily on protease inhibitors because of their ability to rapidly restore CD4 T cell counts and act as a drug barrier (Norris & Rosenberg, 2002). The main obstacle in advancing HIV infection treatment is the extensive genetic variability of the virus. The use of protease inhibitors presents significant challenges in AIDS treatment. These drugs inhibit viral proteases, thereby preventing the cleavage of amino acid chains and formation of proteins necessary for assembling new virus variants.

To develop effective HIV protease inhibitors, it is crucial to identify the cleaved eight-residue peptide accurately. There are 208 possible combinations of 20 amino acids, highlighting the need for a precise and efficient method for predicting HIV protease activity (You et al., 2005).

2. Literature Review

The prediction of HIV-1 protease cleavage sites is crucial for developing effective inhibitors against HIV, and deep convolutional neural networks (CNNs) have been increasingly utilized for this purpose. These models leverage the ability of CNNs to extract complex features from sequence data, enhancing prediction accuracy. Various studies have explored different methodologies, combining CNNs with other machine learning techniques to improve performance. Below, key approaches and findings from recent research are discussed. Multi-View Feature Extraction and Ensemble Learning

A novel approach integrates multi-view feature extraction with a fuzzy rank-based ensemble method, utilizing CNNs for feature extraction. This method combines sequence order effects and physicochemical features, achieving high accuracy and AUC scores, demonstrating its effectiveness in predicting HIV-1 protease cleavage sites (Palmal et al., 2023).The EM-HIV model employs ensemble learning with biased support vector machines and asymmetric bagging to address data imbalance and noise. It uses features from amino acid identities, chemical properties, and coevolutionary patterns, outperforming state-of-the-art models in several evaluation metrics (Hu et al., 2022). Another study uses hybrid descriptors from octapeptide sequences, including bond composition and amino acid binary profiles, with various classifiers. Logistic regression and multi-layer perceptron classifiers showed comparable performance to state-of-the-art models, indicating the potential of hybrid descriptors in improving prediction accuracy (Onah et al., 2022). A hybrid model combines deep CNNs with SVM and genetic algorithms, optimized using metaheuristic algorithms like moth search and dragonfly. This approach enhances the prediction of cleavage sites by fine-tuning activation functions, demonstrating superior performance compared to existing techniques (Kaur & Ghai, 2021). The PU-HIV model introduces a positive-unlabeled learning approach, treating unknown sites as unlabeled rather than negative. This method, using biased SVMs, improves prediction accuracy by reducing bias from false negatives, offering insights into novel inhibitor design (Li et al., 2021).

3. Methodology

Deep Convolutional Neural Networks (DCNNs) have revolutionized computer vision tasks, achieving state-of-the-art performance in object detection, classification, text recognition, and scene understanding (Nguyen et al., 2015). These networks automatically extract hierarchical and translational-invariant spatial features, integrating them with neural network-based classifiers. (Zhou et al., 2016). DCNNs have demonstrated exceptional capabilities in various domains, including image classification, hyperspectral image analysis, and medical pattern recognition (Hu et al., 2015; Khalil et al., 2021; Li et al., 2018). Interestingly, while DCNNs have shown remarkable success, some research suggests that combining supervised and unsupervised deep learning approaches can further improve performance. For instance, stacking DCNN on top of unsupervised layers or replacing DCNN layers with corresponding learnt layers from Convolutional Deep Belief Networks (CDBN) can enhance recognition accuracy and reduce computational costs (Nguyen et al., 2015). Additionally, incorporating recurrent connectivity within convolutional layers, as seen in the Inception Recurrent Convolutional Neural Network (IRCNN), has shown improved performance in object recognition tasks. (Alom et al., 2021).

The suggested model consists of several distinct phases:

- Selecting the data sequence from the information repository
- Data separation and preparation
- Attribute identification method based on 1DWT
- DCNN-based learning and categorization.

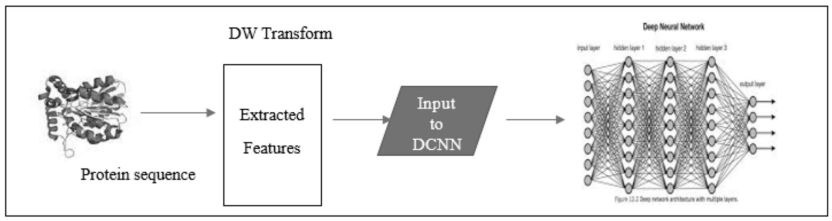

Fig 1: Working Model of DCNN

Gathering information from the domain is extremely difficult, particularly when managing natural information. In this study on cleavage sites, data were collected from the UCI Data Repository. The UCI Repository comprises datasets, domain theories, and data generators utilized by the ML community for the experimental examination of ML calculations. Four datasets were used in the proposed study: Data_set_746, Data_set_1625, Dataset_Schilling, and Dataset_Impens. Each information question comprised two sections: an initial eight-letter alphabetic string representing eight unique amino acids, where "1" and "-1" denote cleavage and non-cleavage spots, respectively, in the octamer. The alphabetic string considered for encoding comprises {B,A,D,N,C,Q,E,H,I,L,M,,K,F,S,P,T,W,V,Y}, each signifying a different AA. Octamer sequence encoding is vital

for understanding the ML methods. Researchers have devised various encoding strategies, such as Ortho-normal Encoding (OE), which consists of a 20-bit vector. However, a drawback of the OE is the loss of useful data. Other strategies incorporate the BLOSUM62, BLOSUM 50 frameworks, and Taylor Venn chart encoding. The proposed system model includes orthonormal encoding, which integrates the inherent and chemical characteristics of each amino acid. Features such as polarity, HI, Hydropathy Index, and proline content were normalized using specific formulas. Various techniques have been incorporated for feature identification, such as KNN, SVM, GP, and ANN; however, the method applied in the proposed model for extracting features is the 1DWT. After extracting the features, the DWT results were used to train the DCNN. Classification involves SL techniques and learning through the properties described in data_input. Classification is performed in two stages: training and testing. The classifier used in this study is the DCNN.

4. Results

This comparative analysis of DCNN and SVM_GA across four datasets (Data746, Data1625, Data Shilling, and Data Impens) reveals notable patterns in their performance, as measured by the Net Predicted Value (NPV). For Data746, DCNN slightly outperforms SVM_GA with an NPV of 88.889 versus 85.714, suggesting a minor advantage in predictive capability for this dataset. However, the difference in performance is minimal. In the case of Data1625, both models exhibit identical performance, each achieving an NPV of 42.857. This indicates that neither model has a clear advantage for this particular dataset, nor both may encounter similar challenges or benefits from its structure. DCNN significantly surpasses SVM_GA when applied to DataShilling, achieving an NPV of 69.369 compared to 48.649. This substantial difference implies that DCNN is particularly well-suited to handle the characteristics or patterns present in DataShilling. Conversely, SVM_GA demonstrates superior performance on Data Impens, with an NPV of 71.795 versus DCNN's 56.41. This suggests that SVM_GA may be better equipped to manage the unique features of Data Impens, possibly due to its capacity to handle intricate decision boundaries more effectively in this instance. The results indicate that neither model consistently outperforms the other across all datasets. DCNN exhibits notable strengths in certain scenarios, particularly with Data746 and Data Shilling, while SVM_GA shows superior performance with Data Impens. The comparable performance on Data1625 further emphasizes that model effectiveness is heavily influenced by the specific characteristics of each dataset. These findings suggest that the selection of an appropriate model should be tailored to the nature of the data and the problem at hand.

Table 1: Negative Predicted Value for Four datasets

Data Set	DCNN	SVM GA
Data 746	88.889	85.714
Data 1625	42.857	42.857
Schilling	69.369	48.649
Impens	56.41	71.795

This comparative analysis employs the Matthews Correlation Coefficient (MCC) to evaluate the effectiveness of DCNN and SVM_GA across four datasets: Data746, Data1625, DataShilling, and DataImpens. In the case of Data746, DCNN exhibited superior performance with an MCC of 0.6814, surpassing SVM_GA's 0.62665. This indicates a more robust positive correlation between predicted and actual values for DCNN. A similar pattern emerged with DataShilling, where DCNN achieved 0.65096, while SVM_GA scored 0.59219. These findings suggest DCNN's greater reliability in capturing relationships within these two datasets. However, both models struggled with Data1625, yielding low MCC values (DCNN: 0.26828, SVM_GA: 0.23761), though DCNN maintained a marginal advantage. Notably, SVM_GA slightly outperformed DCNN for DataImpens, scoring an MCC of 0.55176 compared to DCNN's 0.5142. This implies that SVM_GA might be better equipped to handle the intricacies or characteristics present in DataImpens. In summary, DCNN demonstrates a consistent advantage over SVM_GA in most scenarios, particularly for Data746 and DataShilling, where its higher MCC reflects stronger predictive capabilities. However, SVM_GA's superior performance with DataImpens indicates that model selection should consider the specific attributes of the dataset in question. The comparable performance of both models on Data1625 highlights the difficulties presented by this particular dataset. Consequently, the optimal model choice should take into account the nature of the dataset and the intended predictive outcomes.

Table 2: Mathew Correlation Coefficient for Four datasets

Data Set	DCNN	SVM GA
Data 746	0.6814	0.62665
Data 1625	0.26828	0.23761
Schilling	0.65096	0.59219
Impens	0.5142	0.55176

This comparative analysis utilized the F1 score to evaluate the effectiveness of DCNN and SVM_GA across four datasets: Data746, Data1625, Data Shilling, and Data Impens. For Data746, DCNN exhibited a slight advantage over SVM_GA, achieving an F1 score of 81.481 compared to 78.534. This indicates that DCNN offers a somewhat superior balance of precision and recall in predicting correct outcomes. In the case of Data1625, both models demonstrated nearly identical performance, with DCNN and SVM_GA scoring 96.67 and 96 respectively. This close result suggests that both models are highly effective in predicting values for this particular dataset. Examining DataShilling, both models showed exceptional performance, with DCNN reaching an F1 score of 96.037 and SVM_GA attaining 96.081. The minimal difference between these scores underscores their comparable predictive capabilities for this dataset, indicating that they are almost equally robust when applied to DataShilling. For Data Impens, DCNN achieved a score of 93.496, while SVM_GA scored 92.662. This outcome again shows DCNN outperforming SVM_GA, albeit by a small margin. In summary, DCNN demonstrates a slight advantage over SVM_GA in most scenarios, particularly for Data746 and Data Impens, where it achieves higher F1 scores. However, the models performed nearly identically for Data1625 and DataShilling, with only minor variations in their F1 scores. These findings suggest that DCNN may be more suitable for datasets like Data746 and Data Impens, while both models prove highly effective for Data1625 and DataShilling, where they achieve near-perfect scores. Consequently, the selection between these models depends on the specific requirements and characteristics of the dataset in question.

Table 3: F1 Score for Four datasets

Data Set	DCNN	SVM GA
Data 746	81.481	78.534
Data 1625	96.67	96
Schilling	96.037	96.081
Impens	93.496	92.662

5. Conclusion

This research investigates the current status of predictive models for discovering HIV-1 protease cleavage sites in proteins and proposes a new approach combining feature extraction and classification using machine learning methods, specifically Deep Convolutional Neural Networks (DCNNs). The study compares the performance of DCNN with SVM_GA across four datasets (Data746, Data1625, Data Shilling, and DataImpens) using various evaluation metrics such as Net Predicted Value (NPV), Matthews Correlation Coefficient (MCC), and F1 score. The results indicate that DCNN generally outperforms SVM_GA, particularly in Data746 and Data Shilling, while SVM_GA shows superior performance in Data Impens. However, the performance of both models is comparable in Data1625. The findings suggest that model selection should be tailored to the specific characteristics of the dataset and the problem at hand.

References

1. Briand, L. C., Daly, J., and Wüst, J., "A unified framework for coupling measurement in object oriented systems", IEEE Transactions on Software Engineering, 25, 1, January 1999, pp. 91-121.

2. Maletic, J. I., Collard, M. L., and Marcus, A., "Source Code Files as Structured Documents", in Proceedings 10th IEEE International Workshop[1]Alom, M. Z., Yakopcic, C., Hasan, M., Taha, T. M., & Asari, V. K. (2021). Inception recurrent convolutional neural network for object recognition. Machine Vision and Applications, 32(1). https://doi.org/10.1007/s00138-020-01157-3

3. Buvé, A., Bishikwabo-Nsarhaza, K., & Mutangadura, G. (2002). The spread and effect of HIV-1 infection in sub-Saharan Africa. The Lancet, 359(9322), 2011–2017. https://doi.org/10.1016/s0140-6736(02)08823-2

4. Ghosh, A. K., Osswald, H. L., & Prato, G. (2016). Recent Progress in the Development of HIV-1 Protease Inhibitors for the Treatment of HIV/AIDS. Journal of Medicinal Chemistry, 59(11), 5172–5208. https://doi.org/10.1021/acs.jmedchem.5b01697
5. Gilbert, M. T. P., Rambaut, A., Spira, T. J., Pitchenik, A. E., Worobey, M., & Wlasiuk, G. (2007). The emergence of HIV/AIDS in the Americas and beyond. Proceedings of the National Academy of Sciences, 104(47), 18566–18570.https://doi.org/10.1073/pnas.0705329104
6. Gulnik, S., Erickson, J. W., & Xie, D. (2000). HIV protease: Enzyme function and drug resistance. Vitamins and Hormones,58,213–256.https://doi.org/10.1016/s0083-6729(00)58026-1
7. Hu, W., Li, H., Zhang, F., Wei, L., & Huang, Y. (2015). Deep Convolutional Neural Networks for Hyperspectral Image Classification. Journal of Sensors, 2015(2015), 1–12. https://doi.org/10.1155/2015/258619
8. Khalil, A., El‐Shafai, W., Abd El‐Samie, F. E., Rihan, M., Mahrous, Y., Dessouky, M. I., El‐Rabaie, E. M., Soltan, E., El‐Banby, G. M., Elsherbeeny, Z., Saleeb, A. A., El‐Bendary, M. A. M., El‐Fishawy, A. S., Khalaf, A. A. M., E Ibrahim, F., Haggag, N., Messiha, N. W., Algarni, A. D., Soliman, N. F., … El-Dokany, I. (2021). Efficient anomaly detection from medical signals and images with convolutional neural networks for Internet of medical things (IoMT) systems. International Journal for Numerical Methods in Biomedical Engineering, 38(1). https://doi.org/10.1002/cnm.3530
9. Li, J., Du, Q., Li, Y., Xi, B., Zhao, X., & Hu, J. (2018). Classification of Hyperspectral Imagery Using a New Fully Convolutional Neural Network. IEEE Geoscience and Remote Sensing Letters, 15(2), 292–296. https://doi.org/10.1109/lgrs.2017.2786272
10. Nguyen, K., Fookes, C., & Sridharan, S. (2015). Improving deep convolutional neural networks with unsupervised feature learning. 11, 2270–2274. https://doi.org/10.1109/icip.2015.7351206
11. Norris, P. J., & Rosenberg, E. S. (2002). CD4(+) T helper cells and the role they play in viral control. Journal of Molecular Medicine (Berlin, Germany), 80(7), 397–405. https://doi.org/10.1007/s00109-002-0337-3
12. You, L., Garwicz, D., & RöGnvaldsson, T. (2005). Comprehensive bioinformatic analysis of the specificity of human immunodeficiency virus type 1 protease. Journal of Virology, 79(19), 12477–12486. https://doi.org/10.1128/jvi.79.19.12477-12486.2005
13. Zhou, Y., Xu, F., Jin, Y.-Q., & Wang, H. (2016). Polarimetric SAR Image Classification Using Deep Convolutional Neural Networks. IEEE Geoscience and Remote Sensing Letters, 13(12), 1935–1939. https://doi.org/10.1109/lgrs.2016.2618840
14. Palmal, S., Saha, S., & Tripathy, S. (2023). Integrating Multi-view Feature Extraction and Fuzzy Rank-Based Ensemble for Accurate HIV-1 Protease Cleavage Site Prediction (pp. 480–492). Springer Science+Business Media. https://doi.org/10.1007/978-981-99-8141-0_36
15. Hu, L., Li, Z., Tan, Z., Zhao, C., & Zhou, X. (2022). Effectively predicting HIV-1 protease cleavage sites by using an ensemble learning approach. BMC Bioinformatics, 23(1). https://doi.org/10.1186/s12859-022-04999-y
16. Onah, E. I., Uzor, P. F., Ugwoke, I. C., Eze, J. U., Ugwuanyi, S. T., Chukwudi, I. R., & Ibezim, A. (2022). Prediction of HIV-1 protease cleavage site from octapeptide sequence information using selected classifiers and hybrid descriptors. BMC Bioinformatics, 23(1). https://doi.org/10.1186/s12859-022-05017-x
17. Kaur, N., & Ghai, W. (2021). Performance Analysis of Deep CNN Assisted Optimized HIV-I Protease Cleavage Site Prediction with Hybridized Technique (pp. 529–540). Springer, Singapore. https://doi.org/10.1007/978-981-33-4909-4_40
18. Li, Z., Hu, L., Tang, Z., & Zhao, C. (2021). Predicting HIV-1 Protease Cleavage Sites With Positive-Unlabeled Learning. Frontiers in Genetics, 12(3), 658078. https://doi.org/10.3389/FGENE.2021.658078
19.]Hu, L., Hu, P., Luo, X., Yuan, X., & You, Z.-H. (2020). Incorporating the Coevolving Information of Substrates in Predicting HIV-1 Protease Cleavage Sites. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 17(6), 2017–2028. https://doi.org/10.1109/TCBB.2019.2914208, Prediction of HIV-1 Protease Cleavage Site from Octapeptide Sequence Information using Selected Classifiers and Hybrid Descriptors. (2022). https://doi.org/10.21203/rs.3.rs-1688464/v1
20. Lu, X., Wang, L., & Jiang, Z. (2018). The Application of Deep Learning in the Prediction of HIV-1 Protease Cleavage Site. International Conference on Systems, 1299–1304. https://doi.org/10.1109/ICSAI.2018.8599496
21. Singh, D., Singh, P. K., & Sisodia, D. (2019). Evolutionary based ensemble framework for realizing transfer learning in HIV-1 Protease cleavage sites prediction. Applied Intelligence, 49(4), 1260–1282. https://doi.org/10.1007/S10489-018-1323-Y