A Machine Learning Approach to Classify Carcinomas and Infections According to Viral Integration

MD. Sakib Sami
Electrical & Computer Engineering
North South University
Dhaka, Bangladesh
sakib.sami@northsouth.edu

Tamim Ishrak Sanjid
Electrical & Computer Engineering
North South University
Dhaka, Bangladesh
tamim.sanjid@northsouth.edu

Farhan Ishrak Tahmid
Electrical & Computer Engineering
North South University
Dhaka, Bangladesh
farhan.tahmid05@northsouth.edu

Osmita Monzur Subonty
Electrical & Computer Engineering
North South University
Dhaka, Bangladesh
osmita.subonty@northsouth.edu

Abstract*—Diseases and infections caused by viruses are a major health concern worldwide, often progressing without noticeable symptoms. While some cancers are thought to arise from genetic or lifestyle factors, many are linked to viral infections. Viruses can integrate into the host genome, disrupting tumor suppressor genes, altering gene expression, and causing genomic instability, which can lead to severe diseases, including cancer. This study introduces a machine learning-based approach to classify infections and cancers caused by viral integration automatically. Unlike previous research focused on a single disease, this paper provides a comprehensive framework to identify multiple diseases linked to viral integration sites. The novelty of this work lies in enabling early predictions and improving the chances of detecting carcinomas and infections before they become severe. We used the publicly available ViMIC dataset, initially comprising data on eight viruses and 77 human diseases, totaling 105,642 entities. After cleaning and preprocessing the data, the scope was refined to six viruses and ten diseases, resulting in 32,151 entities. Four feature selection techniques were applied, and the data was trained using nine different machine learning models. Models trained on features selected through variance threshold performed poorly, whereas RFECV (Recursive Feature Elimination with Cross-Validation) yielded the best results. To further enhance performance, hyperparameter tuning was conducted using RandomizedSearchCV and GridSearchCV. Among the models tested, Random Forest and XGBoost Classifier stood out, achieving an impressive accuracy of 92%. The predictions were explained using LIME, an explainable AI framework, ensuring transparency and interpretability of the results. This work fills a significant gap in the literature by offering a robust and scalable framework for classifying diseases based on viral integration, facilitating timely intervention and better patient outcomes.

***Index Terms*—viral integration, machine learning, supervised learning, explainable ai, artificial intelligence, random forest, disease classification, carcinomas, infections**

I. Introduction

Viruses are pathogens (germs or micro-organisms that cause diseases) that can result in temporary viral infections. Most viral infections are not dangerous and tend to cure over time; however, some infections can be long-lasting and extreme [1]. HIV and cancer are among the two extreme cases. According to [2], viral infection is the central cause of cancer for more than **1,400,000** cases annually. **(15-20)%** of all human cancers are caused by oncogenesis (a multi-step process that turns healthy cells into cancerous cells). Epstein–Barr Virus (EBV), Hepatitis B Virus (HBV), Human T-lymphotropic Virus 1 (HTLV1), Human Papillomaviruses (HPV), Hepatitis C Virus (HCV) are widely known among the oncogenic categories [3]. Oncogenic viruses function as a biological process to replicate, which would not be detected by the immune system of the host cell in the early stage. The 2018 Global Cancer Observatory data showed that about **2.2 million** diagnosed cancers were due to oncogenic viruses. Not all infected by the viruses develop cancer, but those with weak immune systems have a higher probability [4].

Additionally, by the end of 2023, an estimated **39.9 million** people were affected by the Human Immunodeficiency Virus (HIV), the virus that causes AIDS [5]. Epstein-Barr virus and HIV have been linked to an increased risk of developing certain childhood cancers. Cancer is considered the second leading cause of death in the United States [6]. Many assume the origin of cancer to be from family genes or lifestyle factors, such as smoking, drinking, and poor diet [7]. However, viral infection can be of major significance here. Possibly, the virus alters a cell in some way. That cell then reproduces an

altered cell, and eventually, these alterations become cancer cells that reproduce more cancerous cells. As a step towards that, this study attempts to classify various types of carcinomas and infections throughout the globe by how viruses integrate themselves into the cells and alter their characteristics. For example, viruses such as HPV and HBV can integrate their genome into their host genome, and this genome integration is considered to be the major mechanism for the carcinogenic effects [8], driving toward the mutation of a certain type of cancer.

A machine learning technique has been integrated to help classify diseases and infections automatically. Cancers or infections from viruses can be unnoticeable, with no early symptoms. The uniqueness of this paper lies in facilitating the chances of early predictions to determine whether a patient has infectious cells progressing toward the cancerous or infectious path. Additionally, from our knowledge, there is no prior research in this field apart from [9], which focused on only one disease, HTLV1, and our approach aims to fill this gap. The major contributions of this work are as follows:

- Publicly available data set, ViMIC [10], was used, which has been preprocessed and cleaned.
- Machine learning models have been applied to classify the ten diseases, and hyperparameter optimizers and feature selection techniques were used to improve performance.
- LIME, an explainable AI framework tool, was used to evaluate the results predicted by the models.

The ten diseases and infections classified in this paper are HIV infection/AIDS, Hepatocellular carcinoma, Adult T cell leukemia/lymphoma, HTLV1 associated myelopathy/tropical spastic paraparesis, Head and neck squamous cell carcinoma, Cervical carcinoma, B-cell lymphoma, Nasopharyngeal carcinoma, Squamous-cell carcinoma and Gastric carcinoma.

The following is the order in which the paper has been organized. Related works are addressed in Section II. Section III provides how the data set has been cleaned, preprocessed, visualized, and algorithms used. Section IV examines the outcomes achieved from the training. Section V discusses about the outcomes. Section VI highlights some limitations of this paper, and finally, Section VII includes the conclusion with a remark on our future work.

II. Literature Review

Most viruses contribute to various human diseases, including cancers. These molecular mechanisms of viral-related diseases involve multiple factors, including viral mutation accumulation and integration of a viral genome into the host DNA [10]. This integration can lead to cell death [11], genomic instability, disease development, and altered gene expression. Researchers have found that there is a clear relationship between HTLV1 integration and disease [12]. HTLV1 preferably integrates into transcriptionally active regions of chromosomes [12]. A study by Schroder et al. took advantage of the published human genome sequence and demonstrated that in vivo, the sites of HIV-1 integration were not random but rather favored specific chromosomal features, such as transcription units (Schroder et al. 2002) [12]. The fraction of oropharyngeal squamous cell carcinomas (OPSCCs) arising from HPV infection has been around **20(%)**. Integration of HPV DNA into the host genome disrupts tumor suppressor genes and promotes oncogene expression [13]. The relationship between AAV and the host remains obscure due partially to the absence of associated pathology [11]. Timely identification of these human-infecting viruses is crucial for understanding their role in carcinogenesis and for implementing appropriate interventions [13], [14].

Although next-generation sequencing has led us to a path to identify viral and host genomic sequences and their role in disease, the application of machine learning to classify diseases based on viral integration sites remains unexplored. Related research essentially focused on detecting viral integration sites [8], viral integration recognition [15], detecting HTLV1 virus integration [9] etc.

From our knowledge, no prior research has classified diseases based on viral integration sites using data from diverse viruses. Our research filled this gap by utilizing the ViMIC database, a curated human disease-related virus mutations database, integration sites, and Cis-effects [10].

III. Methodology and Model Architecture

This approach involves some necessary steps, shown in Figure 1. The dataset was collected from the ViMIC [10], containing viral integration and disease association information across different countries. The collected data was preprocessed, categorical features were encoded, and the numerical features were scaled. Following the split into training **(80%)** and test set **(20%)**, various machine learning models were used to train on the data. Different feature selection methods were applied.

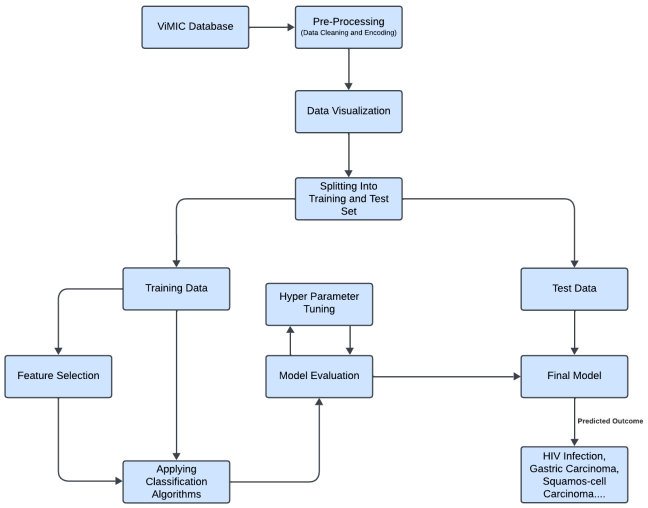


Fig. 1: Workflow

A. Data Collection

The database contains all the necessary information about viral integration sites. The dataset has **1,05,624 entries** of viral integration sites of eight viruses in 77 human diseases obtained from the public domain. The major source of this data in ViMIC was the published literature. The key features for each VIS (Virus Integration Sites) include start breaking point, end breaking point, target gene, chromatin accessibility, histone modification, sum overlaps, viral reference genome, human reference genome, chromosome, and associated disease.

B. Data Cleaning

The files containing data about viral integration sites were imported from ViMIC and then merged into a single data frame. Before merging, the 'Virus Name' and 'Virus Type' columns were created, where every instance of the data set was labeled as DNA (deoxyribonucleic acid) or RNA (ribonucleic acid) type based on the virus name. The dataset was highly imbalanced, and some data were missing. The misspelled tumor was corrected in the samples and labeled 'unknown' as 'unsure.' Some unnecessary symbols were removed. In the VRG column, some viral reference genomes integrated into the host genome were grouped with semicolons, which were split for easy encoding. Some diseases, chromosomes, samples, and VRG were dropped to reduce the class imbalance problem. The numeric features such as histone, sum overlaps, factor, and chromatin accessibility had null values replaced by their corresponding mean value.

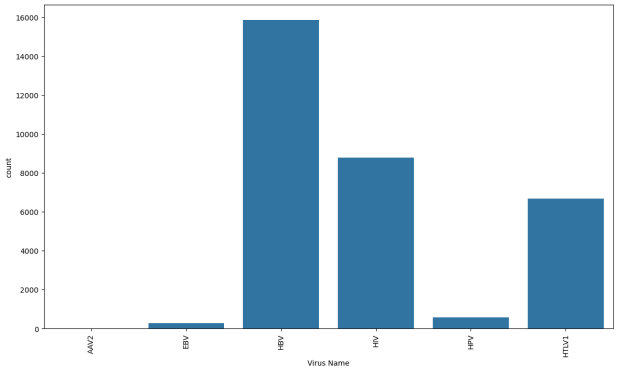
C. Data Visualization

This approach focused on six types of viruses that can cause the corresponding diseases. Adeno-Associated Virus 2 (AAV2), Epstein-Barr Virus (EBV), Hepatitis B Virus (HBV), Human Immunodeficiency Virus (HIV), Human Papillomavirus (HPV), and Human T Lymphotropic Virus type 1 (HTLV1). Figure 2a shows the kinds of viruses and their corresponding counts used to train the models in this paper. These viruses can be RNA (ribonucleic acid) or DNA (deoxyribonucleic acid). According to Figure 2b, **51.9%** of the viruses are DNA, and the rest **48.1%** is RNA, where AAV2, EBV, HBV, HPV are DNA and HIV, HTLV1 is RNA.

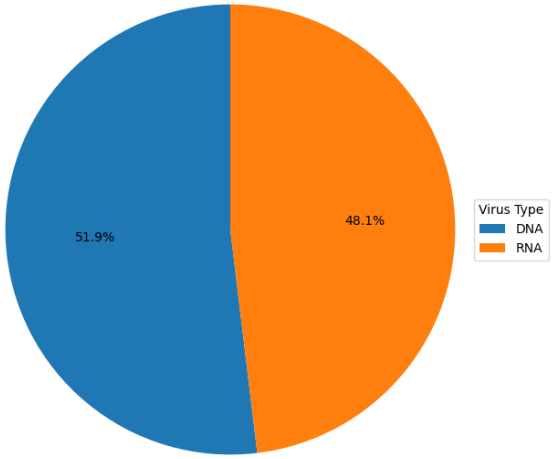
After cleaning and preprocessing the data set [10], the disease counts are shown in Figure 3. According to the figure, Hepatocellular carcinoma is the most counted disease in this data set, having entities more than **15,000**. Each virus may contribute to the formation of the targeted disease, and to better understand the contribution of the viruses towards the disease formation, Figure 4 shows a heatmap. The heatmap highlights the intensified regions that indicate which viruses contribute most to a specific disease. Out of the six, HPV, EBV, and HTLV1 seem to have contributed to more than one virus.

D. Algorithms application

After data cleaning and preprocessing, **32,151** entities were found. The data set was divided as **20%** test set and remaining as the training set. Since the target variable, the



(a) Virus Count



(b) Virus Type

Fig. 2: Types of viruses (a) and their corresponding class (b)

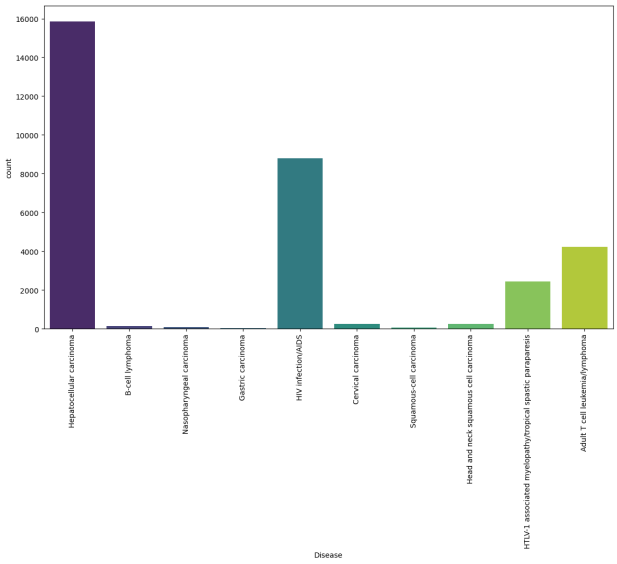


Fig. 3: Disease Count

disease is already labeled in the data set [10], a supervised learning approach was used to train algorithms and create the hypothesis. In this study, a total of nine algorithms were used to train the models using the preprocessed data set, which are Logistic Regression, Gradient Boosted Algorithm, XGBoost Classifier, K-Nearest Neighbor, Random Forest, Decision Tree, Support Vector Machine (both soft and hard) and Naive Bayes (Gaussian). The ZeroR Classifier was used as a benchmark to check whether the nine models are performing as well as expected since the ZeroR classifier is a dummy model. Table I shows the list of all the mentioned models and their accuracy, precision, recall, and f1-score metrics in %, where no feature selection techniques were used and simply the columns with unique identities were dropped.

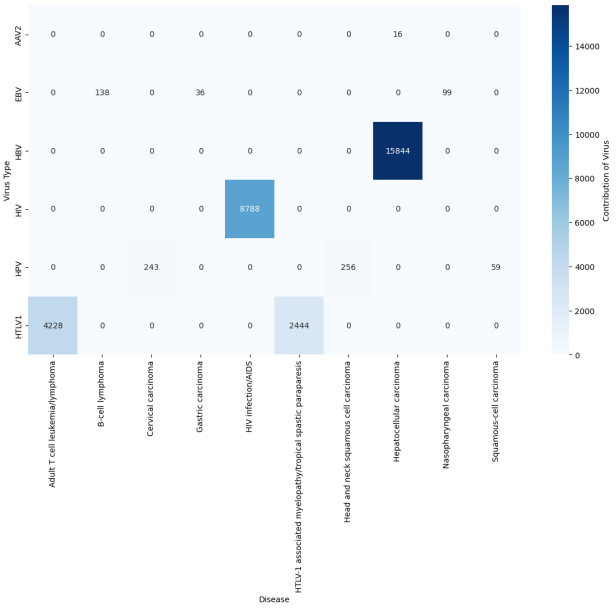


Fig. 4: Heat map of diseases caused by viruses.

Models with high accuracy scores are not sufficient alone. For this study, the recall metrics is significantly important. Recall is the proportion of all actual positives that were classified as positives. The higher the recall, the lower the False Negative (FN) error and the better the model accurately classifies diseases.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}} \tag{1}$$

where:

$$\text{TP : True Positive,}$$
$$\text{FN : False Negative.}$$

Decision Tree [16] is one of the most commonly used classifiers that use the degree of impurity and information gain to select the best possible feature as the root node. The algorithm then splits the data set to make the nodes more homogeneous. One of the core algorithms of the decision tree is Iterative Dichotomiser 3 (ID3). ID3 uses information gain to iteratively select the best attribute at each step, building the tree in a top-notch greedy manner. The degree of impurity of each node is measured using Entropy. A node with zero entropy implies the node being entirely homogeneous; that is, all node instances belong to the same class.

$$\text{E} = \sum_{i=1}^{c} -p_i \log_2 p_i \tag{2}$$

where:

$$\text{c : number of classes}$$
$$\text{p}_i \text{ : probability of i-th category}$$

To determine which attribute is best for the split, ID3 calculates the Information Gain by subtracting the parent node's entropy from its children's weighted entropy. The more positive the information gained from an attribute, the more likely it would be chosen.

$$\text{Information Gain} = Entropy(p) - \sum_{i=1}^{k} \frac{n_i}{n} Entropy(i) \tag{3}$$

where:

$$\text{p : parent node}$$
$$\text{k : children (node)}$$
$$\text{n : number of records in parent node}$$
$$\text{n}_i \text{ : number of records in child node i}$$

According to Table I, the decision tree classifier has an accuracy metric of **91%**.

Random Forest [17], an ensemble classifier where several decision trees are constructed from the data set and are used to predict the output either by majority voting or through averaging the output made by each decision tree. Random forest utilizes random feature selection, a technique where each decision tree classifier is assigned a subset of the original feature set. It also supports bagging, a method to sample the data set with repetition among the decision tree classifiers. This randomness introduces variability among the individual classifiers, reducing the risk of over-fitting and enhancing the performance. In this approach, the random forest classifier has been trained by setting both oob_score positive and negative. However, there was no difference in the metrics, where accuracy remained at **92%** under the RFECV technique.

XGBoost Classifier [18], an ensemble classifier, implements gradient-boosted decision trees. In this algorithm, the base models depend on one another, and weights are assigned to each instance and fed into the decision tree. The weights of the instances predicted wrong by the first decision tree are increased so that the next decision tree, trained on this data set, can focus more on the error. This classifier can handle class imbalance very well, and since the data set used in this study is partially imbalanced even after cleaning and preprocessing,

the model still performed very well, having an accuracy score of **92%** according to Table I.

IV. Result Analysis

Algorithms	Accuracy	Precision	Recall	F1 Score
Logistic Regression	91%	85%	91%	87%
Decision Tree	91%	91%	91%	91%
Random Forest	92%	91%	92%	91%
Naive Bayes	91%	90%	91%	89%
XGBoost Classifier	92%	91%	92%	91%
KNN	90%	90%	90%	90%
Support Vector Machine	91%	86%	91%	88%
Voting Classifier (Hard)	92%	91%	92%	90%
Voting Classifier (Soft)	92%	92%	92%	91%
ZeroR Classifier	49%	24%	49%	32%

TABLE I: Performance metrics for different algorithms on Dataset - ViMIC [10] without feature selection techniques.

This section highlights the results from the proposed models used in this study. Table I shows all of the metrics used to evaluate each model, which includes accuracy, precision and recall, and f1-score. Equation 1, Equation 4, Equation 5 and Equation 6 were used to calculate the corresponding metrics where TP, TN, FP, and FN are true positive, true negative, false positive, and false negative.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}} \tag{4}$$

$$\text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{5}$$

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}} \tag{6}$$

Based on the type of problem we are solving, precision alone is not sufficient to judge a model, and this study also focused on the recall metric for model evaluation, as mentioned earlier.

A. Performance of different algorithms using various feature selection techniques

Four feature selection techniques were applied to reduce over-fitting and training time and improve accuracy: variance threshold, recursive feature elimination, information gain, and Pearson correlation to find valuable features of the data set. The variance threshold eliminates features with low variance or constant features. Pearson's correlation eliminates highly correlated features, as they indicate duplicate features.

Algorithms	Feature Selection Method	RandomizedSearchCV	GridSearchCV
Decision Tree	Information Gain	92%	92%
	Recursive Feature Elimination	92%	92%
	Pearson Correlation Coefficient	92%	92%
Random Forest	Information Gain	92%	92%
	Recursive Feature Elimination	92%	92%
	Pearson Correlation Coefficient	92%	92%
XGBoost Classifier	Information Gain	92%	92%
	Recursive Feature Elimination	92%	92%
	Pearson Correlation Coefficient	92%	92%

TABLE II: Accuracy of algorithms after hyperparameter tuning.

$$\mathbf{r} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}} \tag{7}$$

where:

r : pearson correlation coefficient

x_i : values of the x-variable in a sample

$\bar{x}$: mean of the values of the x-variable

y_i : values of the y-variable in a sample

$\bar{y}$: mean of the values of the y-variable

Information gain or mutual information is used to measure how much information a feature contributes to making a correct prediction of a target based on the absence or presence of the feature. In simple terms, it measures the dependency between the variables. If two random variables are independent, the value will equal zero. If the value is higher, then there is a higher dependency.

In recursive feature elimination, weights are assigned to features, and features are selected recursively, considering smaller and smaller sets of features. An estimator is trained on the initial set of features, and the importance of each feature is obtained through any specific attribute. Then, the least important features are pruned from the current set of features. The process recursively repeats on the pruned set until the desired number of features is obtained. In recursive feature elimination with cross-validation, each feature is ranked with recursive feature elimination, and the cross-validation technique is used to select the best number of features. Table III lists the performance of all the models for the four feature selection techniques.

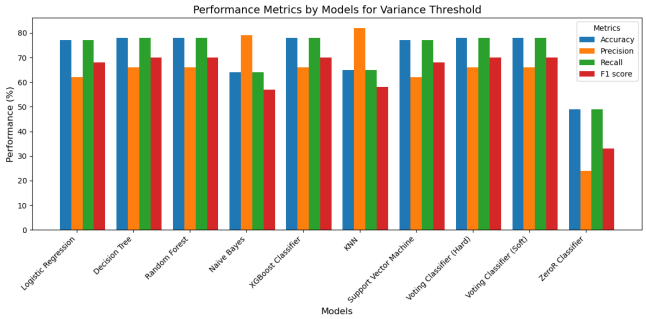
The highest accuracy was obtained by using RFECV in Random Forest and XGBoost, where the optimal number of features was 12. Except for the variance threshold, the remaining feature selection methods were very close to RFECV. In addition to Table III, Figure 5 shows the performance of all the models over the four feature selection technique where variance threshold method performed very poorly.

B. Accuracy of algorithms after using hyperparameter optimizers

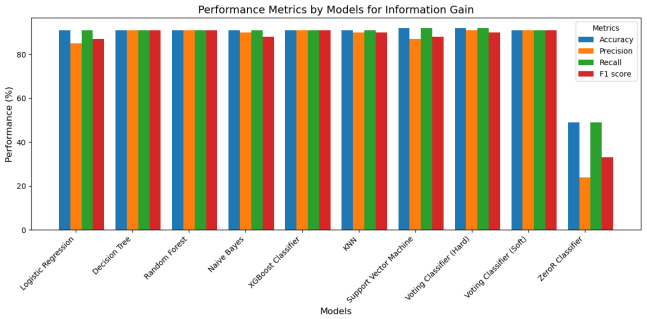
This study also focused on tuning the hyperparameters to improve the models further. Table II shows the list of models on which the hyperparameter optimization was applied. The model's accuracy for some feature selection techniques did improve by **1%**, but, overall, there was no significant improvement through tuning the parameters. The maximum accuracy reached is **92%**, also seen through normal training and evaluation based on Table III.

C. Using LIME to understand the predictions made

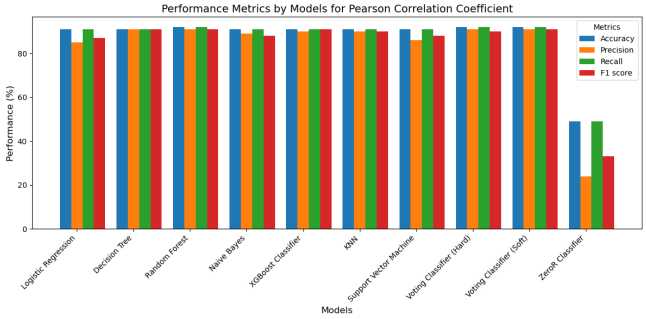
LIME is an explainable AI used to understand the predictions made by the model. LIME was used on the Random Forest Classifier and trained on the features selected by the RFECV technique, as it gave the highest accuracy and



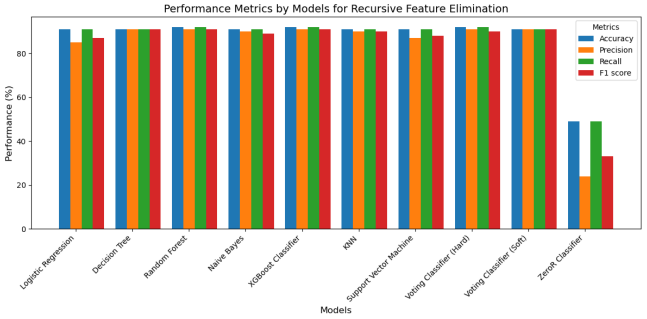
(a) With Variance Threshold



(b) With Information Gain



(c) With Pearson Correlation



(d) With Recursive Feature Elimination

Fig. 5: Visual representation of the algorithmic performance for the four feature selection techniques.

better recall. Figure 6 shows the LIME explanations for three diseases, applied in the test set to understand the reasons why the Random Forest Classifier made this prediction for three specific instances of the test set. Green bars indicate features that push the model's prediction toward the predicted class, and red bars indicate those that push the prediction away from the predicated class.

V. Discussion

Despite the majority of the models used having an accuracy close to 90% or more, according to Table I and Table III, the tree-based models (Decision tree, Random Forest and XGBoost Classifier) have an overall higher recall. They were properly able to classify the correct disease by making minor errors compared to other models. Furthermore, compared to the work done in [9], which is limited to only HTLV1, this study focused on a wider variety of diseases caused by six categories of viruses. Random Forest and XGBoost classifiers have been found to give the best result with high accuracy and recall. Since they are both ensemble classifiers that combine the strength of multiple base classifiers, they are likely to have provided better results.

VI. Limitation

Unlike previous research, this approach focused on a more broader set of diseases and viruses; however, this work has faced some number of difficulties:

- Due to data scarcity, even after cleaning and preprocessing, class imbalance challenges still remained.

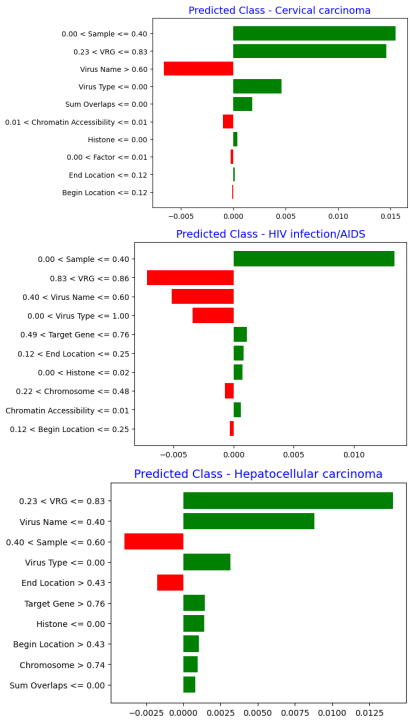


Fig. 6: LIME explanation for HIV, Cervical and Hepatocellular carcinoma.

- According to the description of [10], the used data in this research from ViMIC included data from VISDB

Algorithms	Feature Selection Method	Accuracy	Precision	Recall	F1 score
Logistic Regression	Information Gain	91%	85%	91%	87%
	Recursive Feature Elimination	91%	85%	91%	87%
	Variance Threshold	77%	62%	77%	68%
	Pearson Correlation Coefficient	91%	85%	91%	87%
Decision Tree	Information Gain	91%	91%	91%	91%
	Recursive Feature Elimination	91%	91%	91%	91%
	Variance Threshold	78%	66%	78%	70%
	Pearson Correlation Coefficient	91%	91%	91%	91%
Random Forest	Information Gain	91%	91%	91%	91%
	Recursive Feature Elimination	**92%**	91%	**92%**	91%
	Variance Threshold	78%	66%	78%	70%
	Pearson Correlation Coefficient	92%	91%	92%	91%
Naive Bayes	Information Gain	91%	90%	91%	88%
	Recursive Feature Elimination	91%	90%	91%	89%
	Variance Threshold	64%	79%	64%	57%
	Pearson Correlation Coefficient	91%	89%	91%	88%
XGBoost Classifier	Information Gain	91%	91%	91%	91%
	Recursive Feature Elimination	**92%**	91%	**92%**	91%
	Variance Threshold	78%	66%	78%	70%
	Pearson Correlation Coefficient	91%	90%	91%	91%
KNN	Information Gain	91%	90%	91%	90%
	Recursive Feature Elimination	91%	90%	91%	90%
	Variance Threshold	65%	82%	65%	58%
	Pearson Correlation Coefficient	91%	90%	91%	90%
Support Vector Machine	Information Gain	92%	87%	92%	88%
	Recursive Feature Elimination	91%	87%	91%	88%
	Variance Threshold	77%	62%	77%	68%
	Pearson Correlation Coefficient	91%	86%	91%	88%
Voting Classifier (Hard)	Information Gain	92%	91%	92%	90%
	Recursive Feature Elimination	92%	91%	92%	90%
	Variance Threshold	78%	66%	78%	70%
	Pearson Correlation Coefficient	92%	91%	92%	90%
Voting Classifier (Soft)	Information Gain	91%	91%	91%	91%
	Recursive Feature Elimination	91%	91%	91%	91%
	Variance Threshold	78%	66%	78%	70%
	Pearson Correlation Coefficient	92%	91%	92%	91%
ZeroR Classifier	Information Gain	49%	24%	49%	33%
	Recursive Feature Elimination	49%	24%	49%	33%
	Variance Threshold	49%	24%	49%	33%
	Pearson Correlation Coefficient	49%	24%	49%	33%

TABLE III: Performance metrics for different algorithms with feature selection techniques.

[19], creating challenges in externally validating the performance of the obtained models.

- Processed data had adequate entries of HIV, allowing this study to work around this sole infection, ignoring others due to data scarcity.

VII. Conclusion

This paper focused on detecting ten classes of diseases and infections based on six categories of viruses through the use of machine learning models. A publicly available data set was used from ViMIC [10], which had been preprocessed and cleaned. Next, four feature selection methods were applied to determine which features would provide the best result, and recursive feature elimination showed dominance. Finally, the split data set in the ratio of **20%** as a test case was trained on nine machine learning models. Not all performed well, as accuracy alone won't suffice, recall must be considered as well; however, XGBoost Classifier and Random Forest gave the best result with minimal error. Hyperparameter optimization on three specific models was applied to enhance the performance, where no further improvement in the prediction results was seen. LIME was used to better understand the predictions made by the model on the test set. In the future, the data set could be made more diverse and balanced by collecting more instances of diseases of low count. Other feature selection techniques and hyperparameter optimizations such as the Fisher test, Chi-Square, Manual Uniqueness, etc., and Bayesian Optimization, Generic Algorithms, and Optuna can improve the predicted results further.

References

[1] L. Fayed, "How some viruses cause cancer," 11 2010.

[2] J. T. Schiller and D. R. Lowy, "An introduction to virus infections and human cancer," *Viruses and Human Cancer*, vol. 217, pp. 1–11, 11 2020.

[3] N. A. Krump and J. You, "Molecular mechanisms of viral oncogenesis in humans," *Nature Reviews Microbiology*, vol. 16, pp. 684–698, 08 2018.

[4] G. Ameya and D. J. Birri, "The molecular mechanisms of virus-induced human cancers," *Microbial Pathogenesis*, vol. 183, p. 106292, 10 2023.

[5] W. H. Organization, "Hiv data and statistics," 2024.

[6] R. L. Siegel, A. N. Giaquinto, and A. Jemal, "Cancer statistics, 2024," *CA: A Cancer Journal for Clinicians*, vol. 74, 01 2024.

[7] S. H. Care, "Stanford health care," 2014.

[8] Y. Xia, Y. Liu, M. Deng, and R. Xi, "Detecting virus integration sites based on multiple related sequencing data by virtect," *BMC Medical Genomics*, vol. 12, 01 2019.

[9] H. Xu, J. Jia, H.-H. Jeong, and Z. Zhao, "Deep learning for detecting and elucidating human t-cell leukemia virus type 1 integration in the human genome," *Patterns*, vol. 4, p. 100674, 02 2023.
[10] Y. Wang, Y. Tong, Z. Zhang, R. Zheng, D. Huang, J. Yang, H. Zong, F. Tan, Y. Xie, H. Huang, and X. Zhang, "Vimic: a database of human disease-related virus mutations, integration sites and ¡i¿cis¡/i¿-effects," *Nucleic Acids Research*, vol. 50, pp. D918–D927, 09 2021.
[11] S. Desfarges and A. Ciuffi, "Viral integration and consequences on host gene expression," *Viruses: Essential Agents of Life*, p. 147–175, 09 2012.
[12] S. Jones, "Integration hotspots for disease?," *Nature Reviews Microbiology*, vol. 6, pp. 333–333, 05 2008.
[13] R. Ye, A. Wang, B. Bu, P. Luo, W. Deng, X. Zhang, and S. Yin, "Viral oncogenes, viruses, and cancer: a third-generation sequencing perspective on viral integration into the human genome," *Frontiers in Oncology*, vol. 13, 12 2023.
[14] S. Kumar, S. K. Razzaq, A. D. Vo, M. Gautam, and H. Li, "Identifying fusion transcripts using next generation sequencing," *Wiley Interdisciplinary Reviews: RNA*, vol. 7, pp. 811–823, 08 2016.
[15] D. L. Cameron, N. Jacobs, P. Roepman, P. Priestley, E. Cuppen, and A. T. Papenfuss, "Virusbreakend: Viral integration recognition using single breakends," *Bioinformatics*, vol. 37, pp. 3115–3119, 05 2021.
[16] J. R. Quinlan, "Induction of decision trees," *Machine Learning*, vol. 1, pp. 81–106, 03 1986.
[17] L. Breiman, "Random forests," *Machine Learning*, vol. 45, pp. 5–32, 2001.
[18] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system."
[19] D. Tang, "Visdb - homepage," 2019.