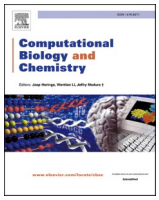
Computational Biology and Chemistry 101 (2022) 107766

Contents lists available at ScienceDirect

Computational Biology and Chemistry

journal homepage: www.elsevier.com/locate/cbac



EBOLApred: A machine learning-based web application for predicting cell entry inhibitors of the Ebola virus

Joseph Adams [a,b], Kwasi Agyenkwa-Mawuli [a,c], Odame Agyapong [a], Michael D. Wilson [b,d], Samuel K. Kwofie [a,c,*]

[a] Department of Biomedical Engineering, School of Engineering Sciences, College of Basic and Applied Sciences, University of Ghana, PMB LG 77, Legon, Accra LG 77, Ghana
[b] Department of Parasitology, Noguchi Memorial Institute for Medical Research (NMIMR), College of Health Sciences (CHS), University of Ghana, P.O. Box LG 581, Legon, Accra LG 581, Ghana
[c] West African Centre for Cell Biology of Infectious Pathogens, Department of Biochemistry, Cell and Molecular Biology, College of Basic and Applied Sciences, University of Ghana, Accra LG 54, Ghana
[d] Department of Medicine, Loyola University Medical Center, Maywood, IL 60153, USA

ARTICLE INFO

Keywords:
Ebola virus protein
Machine learning
Inhibitors
Support vector machine
Random forest
Logistic regression

ABSTRACT

Ebola virus disease (EVD) is a highly virulent and often lethal illness that affects humans through contact with the body fluid of infected persons. Glycoprotein and matrix protein VP40 play essential roles in the virus life cycle within the host. Whilst glycoprotein mediates the entry and fusion of the virus with the host cell membrane, VP40 is also responsible for viral particle assembly and budding. This study aimed at developing machine learning models to predict small molecules as possible anti-Ebola virus compounds capable of inhibiting the activities of GP and VP40 using Ebola virus (EBOV) cell entry inhibitors from the PubChem database as training data. Predictive models were developed using five algorithms comprising random forest (RF), support vector machine (SVM), naïve Bayes (NB), k-nearest neighbor (kNN), and logistic regression (LR). The models were evaluated using a 10-fold cross-validation technique and the algorithm with the best performance was the random forest model with an accuracy of 89 %, an F1 score of 0.9, and a receiver operating characteristic curve (ROC curve) showing the area under the curve (AUC) score of 0.95. LR and SVM models also showed plausible performances with overall accuracy values of 0.84 and 0.86, respectively. The models, RF, LR, and SVM were deployed as a web server known as EBOLApred accessible via http://197.255.126.13:8000/.

1. Introduction

Ebola virus disease (EVD) is a debilitating and often deadly disease caused by the Ebola virus (EBOV) from the filoviridae family (Emanuel et al., 2018; Zawilińska and Kosz-Vnenchak, 2014; Salata et al., 2019). The disease is transmitted from animals to humans through the handling of bushmeat and contact with infected animals including fruit bats, chimpanzees, and forest antelopes (Koch et al., 2020). It is transmitted from person to person via direct contact with body fluids or blood from someone who has died or is sick from the disease (Jacob et al., 2020). Objects contaminated with blood from infected persons can also be a source of transmission (Osterholm et al., 2015). The spread of the disease through direct contact can be via broken skin or mucous membrane in the eyes, nose or mouth (Jacob et al., 2020). Ebola is spread to others only after the manifestation of signs and symptoms. From day 2–21 after being in contact with the virus, individuals show symptoms of fever, aches and pains including severe headache, as well as gastrointestinal symptoms including diarrhea and vomiting (Rajak et al., 2015). Infection with the Ebola virus can be characterized by abnormal blood pressure, other symptoms might also include organ dysfunction syndrome (Leligdowicz et al., 2016). Since the emergence of the Ebola virus disease, more than 33,000 infections in humans have been recorded

Abbreviations: EVD, Ebola virus disease; EBOV, Ebola virus; GP, glycoprotein; VP40, Ebola virus protein 40; AID, assay Identifier; ML, machine Learning; RF, random forest; SVM, support vector machine; NB, Naïve bayes; kNN, k-nearest neighbor; LR, logistic regression; RFE, recursive feature elimination; TP, true positives; FP, false positives; TN, true negatives; FN, false negatives; ROC, receiver operating characteristic; AUC, area under the curve; AD, applicability domain.

* Corresponding author at: Department of Biomedical Engineering, School of Engineering Sciences, College of Basic and Applied Sciences, University of Ghana, PMB LG 77, Legon, Accra LG 77, Ghana.
E-mail address: skkwofie@ug.edu.gh (S.K. Kwofie).

https://doi.org/10.1016/j.compbiolchem.2022.107766
Received 18 March 2022; Received in revised form 10 August 2022; Accepted 29 August 2022
Available online 2 September 2022

with over 40 % death cases (Jacob et al., 2020).

The Ebola virus is a single-stranded, negative-sense RNA virus that forms a threadlike shape with an approximate length of about 970 nm and a uniform diameter of 80 nm (Wan et al., 2017). The virus is tubular and is composed of a viral envelope, matrix, and nucleocapsid components (Johnson et al., 2006). The genome of EBOV comprises nucleoprotein (NP), the viral protein 24 (VP24), polymerase cofactor (VP35), matrix protein (VP40), surface glycoprotein (GP), transcription factor (VP30), and RNA-dependent RNA-polymerase (L) (Qureshi, 2016). VP40 is the main matrix protein and the most abundant protein of the virus (Bornholdt et al., 2013), and it plays a critical role in the virus's life cycle and survival in the host cell. It is responsible for the viral particle assembling, budding, and mediating the exit of the virus from the host cell (Kuhn et al., 2020; Madara et al., 2015). The glycoprotein present at the viral surface is responsible for binding the virus to the host cells and also mediates the fusion of the virus to the host cell membrane (Lee et al., 2008; Jain et al., 2021). Other than supportive therapy, treatments for EVD patients are limited to the use of antibody therapies (Krishnasamy and Saikumar, 2015). Two drug treatments comprising inmazeb and Ebanga block the virus from the host cell receptors. Inmazeb is made of three monoclonal antibodies (atoltivimab, maftivimab and odesivimab), whilst Ebanga is a human monoclonal antibody (ansuvimab) (Lee, 2021). Therapeutic approaches have been investigated so far with varied challenges including remdesivir (GS-5734) (Warren et al., 2016), ZMapp (Davey et al., 2016) and favipiravir (Sissoko et al., 2016).

Efforts are geared towards unraveling new biotherapeutic molecules and vaccines using advanced interventions (Dhama et al., 2018; Schuler et al., 2017; Hansen et al., 2021; Tompa et al., 2021; Madrid et al., 2015). The use of experimental techniques to identify small molecule inhibitors of disease targets presents a tedious, time consuming and a costly approach. Computational methods including molecular docking techniques and machine learning approaches facilitates rapid elucidation of compounds as potential leads (Sliwoski et al., 2013; Kwofie et al., 2019a, 2021; Darko et al., 2021; Asiedu et al., 2021). A pharmacophore-based approach was also used to identify potential inhibitors for EVD (Sankar et al., 2021). Qualitative structure-activity relationship (QSAR) model was used to screen 17 million compounds of which 104 hits were obtained (Capuzzi et al., 2018). Similarly, natural products were identified as potential EBOV protein inhibitors (Kwofie et al., 2019a; Darko et al., 2021). Structural insights pertaining to the matrix protein VP40 from Ebola virus were integrated with biological activity predictions, molecular docking, and dynamics simulations to identify plausible inhibitors via screening plethora of compound libraries (Alam El-Din et al., 2016; Odhar et al., 2019; Khan et al., 2021; Tamilvanan and Hopper, 2013; Nagarajan et al., 2019).

Machine learning Bayesian-based models trained using the EBOV replication and viral pseudotype entry assay datasets were used to screen the MicroSource library to identify inhibitors tilorone, pyronaridine and quinacrine (Ekins et al., 2015). The mechanisms and targets for these inhibitors were corroborated using both *in vitro* and *in vivo* techniques (Lane and Ekins, 2020). Machine learning algorithms have been developed to identify biotherapeutic molecules and these were deployed as web-based applications (Toussi et al., 2021; Agyapong et al., 2021; Sandhu et al., 2021). An anti-Ebola web server was developed to predict anti-EBOV compounds using support vector machine, random forest and artificial neural network models (Rajput and Kumar, 2021). The ML models were trained with anti-Ebola molecules housed in the epidemic and pandemic virus drug and chemical database DrugRepV (Rajput et al., 2021). The models predict inhibitory concentrations with no specific reference to a particular EBOV receptor. ML models trained to predict inhibition against specific inhibitors using deep learning, support vector regression and random forest models have been developed (Espinoza et al., 2021; Bhagwati and Siddiqi, 2020). Likewise, several ML predictive models for SARS-CoV-2 using *in vitro* inhibition data have been developed which are not receptor-specific (Gawriljuk et al., 2021).

Contrary, receptor-specific ML predictive models have been developed for hepatitis C virus NS5B inhibitors using k-nearest neighbor, multi-layer perceptron, partial least squares, random forest and support vectors machine, and implemented as a web server known as StackHCV (Malik et al., 2021). Also, the recursive partition (RP) and naive Bayes (NB) models were trained as classifiers to prioritize HIV-1 Integrase inhibitors (Zhou et al., 2021). Predictive models for identifying inhibitors were also developed for SARS-CoV-2 3C-like protease using ML techniques including convolutional neural network (CNN) (Haneczok and Delijewski, 2021; Kumari and Subbarao, 2021), whilst supervised support vector machine was implemented for elucidating inhibitors of SARS-CoV-2 main protease (Mekni et al., 2021).

In this study, supervised machine learning-based predictive models were developed specifically for EBOV entry inhibitors using random forest (RF), support vector machine (SVM), naïve bayes (NB), k-nearest neighbor (kNN) and logistic regression (LR) approaches. The models were trained with entry inhibitors of Ebola virus-like particles (VLPs) containing the matrix protein VP40 and glycoprotein (GP). The top performing predictive models were implemented as a webserver known as EBOLApred, accessible freely via http://197.255.126.13:8000/.

2. Materials and methods

2.1. Dataset acquisition

The dataset employed for the classification of the models was obtained from PubChem (Wang et al., 2009) with AID 1117304 (Kouznetsova et al., 2014), and consists of 2528 compounds generated from quantitative high-throughput screen (qHTS). The qHTS assay was meant to identify entry inhibitors of Ebola virus-like particles (VLPs) containing the matrix protein VP40 and glycoprotein (GP). Classification of compounds was based on activity score defined by the capability of viral entry inhibition. From the PubChem bioassay AID 1117304 record, the activity of compounds was computed from efficacy, AC_{50} (Tendong et al., 2020) and curve class. AC_{50}, or the dose that generates half-maximum response, is frequently used to summarize compound potencies (Shockley, 2016). Thus, for a given compound, the score was computed using the formulae:

$$\text{Activity score} = [100^{*}((\max(AC_{50})\text{-}AC_{50}(j)) / (\max(AC_{50})\text{-}\min(AC_{50}))) + 100^{*}(1\text{-}(\max(\text{efficacy}) - \text{efficacy}(j)) / (\max(\text{efficacy})\text{-}\min(\text{efficacy})))] / 2 \quad (1)$$

From PubChem (Wang et al., 2009), molecules with activity scores >80 % were classed active, with a total of 596 non-redundant compounds labelled active and 1376 classed inactive.

2.2. Descriptor generation and data pre-processing

PaDEL-descriptor was used to calculate a total of 1445 (1d and 2d) descriptors and 2608 fingerprints (Yap, 2011) using canonical SMILES of the compounds as input. Prior to descriptor calculations, salts were removed from the compounds and the nitro groups were standardized. After descriptor calculations, the dataset was split into a training data, and an external test data in a ratio of 3:1, respectively. For the training data, mean imputation was employed to deal with missing values (Donders et al., 2006; Jerez et al., 2010). The active compounds from the training data formed less than 35 % of the total dataset, thus minority oversampling (Koivu et al., 2020; Perez-Ortiz et al., 2016) module from imbalanced-learn python package (LemaîtreLemaître et al., 2017) was implemented in python (version 3.6) to scale up the active compounds to the level of the inactive (Kumar et al., 2021). The sampling methods tends to create synthetic instances of the minority class to be used for model training (Barua et al., 2011). Synthetic minority oversampling technique (SMOTE) was applied to create synthetic examples of the active compounds. For each active compound, SMOTE works by creating examples of k nearest minority class neighbor (Elreedy and

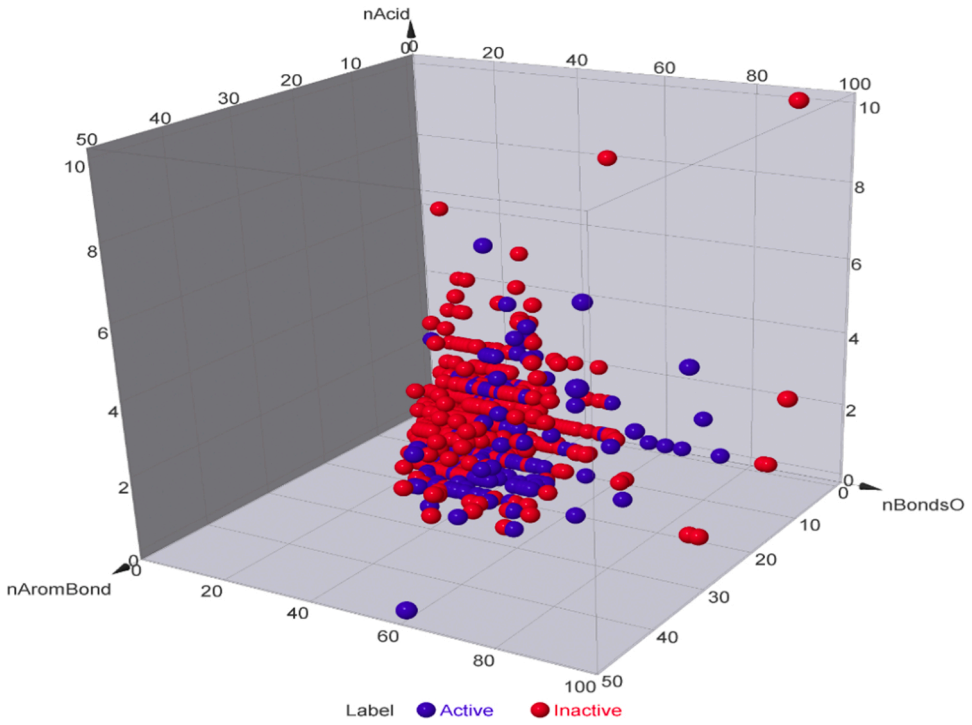


Fig. 1. A 3D scatter plot of bioassay data distribution showing how the active and inactive compounds are related in terms of the numbers of aromatic bonds, oxygen bonds and acids in their chemical structures.

Atiya, 2019). The synthetic example is generated by taking the difference of the feature vector and its nearest neighbor. The difference is then multiplied by a random value between 0 and 1 and the result is added to the vector feature of the minority compound under consideration (Chawla et al., 2011). The technique was used to rectify the class imbalance within the datasets. To further remove redundant features in the training data, dimensionality reduction was performed to reduce the size of the descriptors (Velliangiri et al., 2019), so features with variance less than 0.1 were removed. With class label given to each of the active and inactive compound, *mannwhitneyu* function from scipy library (Virtanen et al., 2020) was used to evaluate the statistical difference between each of the features and the labels. This was based on the p values, where descriptors that had p values >0.5 were deemed to show no statistical difference and were removed from the rest of the descriptors. Recursive feature elimination (RFE) (Chen et al., 2018) was also employed to rank the features selecting the top 612 descriptors as input features for the classification models. The use of raw values after molecular descriptor computation as input to a machine learning model might render the model biased to features with high entries (Ahsan et al., 2021), thus, data scaling was performed for the training data using Standard Scaler from scikit-learn library (Fabianpedregosa et al., 2011) before using them as input features. For each feature, the mean is subtracted, and the result divided by the standard deviation. The standardized data was fitted to develop the predictive models.

$$\text{Standard Scaler(X)} = \frac{X(i) - mean}{std} \quad (2)$$

where i represents each value in the feature X.

2.3. Development of machine learning models

The training dataset was used to build five models from five different machine learning algorithms of which the best performing model based on the classification metrics was selected. The models were built on 70 % training data and 30 % test data from the dataset which had been preprocessed. The five predictive models developed comprised k-nearest neighbors (k-NN), Gaussian Naïve Bayes (GaussianNB), Support vector machine (SVM), Random Forest classifier (RF) and Logistic regression (LR).

The k-nearest neighbor, support vector machine, gaussian naïve bayes, random forest and logistic regression were constructed to comparatively study their performance on the datasets. Each of the classifiers were first optimized to select the best hyperparameter that gave the highest accuracy before being compared to the other optimized classifiers. The kNN was constructed with varying integers of 'k'. Nearest neighbors of 3, 5 and 7 were used in optimizing the model with k = 3 being the most accurate amongst them. SVM model was optimized with radial biased kernel function, regularization parameter C ranging from 0.1 to 2 with a step of 0.1 and gamma values ranging from 0.1 to 1 with a step size of 0.1. Using grid search method, the best fit was obtained for the SVM model with C = 1 and gamma = 0.1. Grid search was once again used to obtain the best fitting hyperparameters for the random forest model. The search produced a random forest model with max_depth of 8 and n_estimators of 120. Default hyperparameter settings for both gaussian naïve bayes and logistic regression models from scikit-learn library were maintained and used for the cross-validation.

2.4. Model validation

The effectiveness of the optimized models was accessed using 10-fold cross-validation (Tougui et al., 2021). The cross-validation technique produces a reliable estimate of the model performance on an unseen data. It works by dividing the training data into k number of groups. The model is then trained on k-1 folds with one reserved for testing (Berrar, 2018). The technique was used to compare the models based on five classification metrics consisting of accuracy, balanced accuracy, precision, recall and F1 score. True positives (TP), false positives (FP), true negatives (TN) and false negatives (FN) forms the confusion matrix on which the predictive metrics were computed (Jiao and Du, 2016; Seliya et al., 2009). The classification accuracy of a predictive model defines the ratio of correct predictions made by the model to the total number of predictions.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \quad (3)$$

The use of SMOTE provided an equal number of active and inactive data for training the ML algorithms. As such the use of accuracy provides a good evaluation metrics and a basis for model comparison (Wei and Dunbrack, 2013). Precision provides the ratio of correctly predicted positives to the total positive observations made by the model (Velliangiri et al., 2019). Recall or sensitivity indicate the ratio of true positive predictions made by the model to the total number of positives observations in the datasets. F1 score provides a statistical means of merging precision and recall (Virtanen et al., 2020). Balanced accuracy defines the average of sensitivity and specificity. It is often employed to access the performance of a model trained on an imbalanced data (García et al., 2009). Based on these metrics, the performances of the ML models were assessed.

$$\text{Precision} = \frac{TP}{TP + FP} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP + FN} \quad (5)$$

$$\text{F1} = 2x\frac{Precision * Recall}{Precision + Recall} \quad (6)$$

$$\text{Balanced } accuracy = \frac{Sensitivity + Specificity}{2} \quad (7)$$

2.5. Model deployment

The three top performing models were implemented as a web application with the front end designed with html and css. The backend of the server was built with Django 4.0.3 (Juneau et al., 2010), python (Taneja and Gupta, 2014) and JavaScript (bin Uzayr et al., 2019).

3. Result and discussion

3.1. Data acquisition and processing

The bioactive dataset acquired from PubChem consisted of an imbalanced data for which the actives formed approximately one third of the total datasets. As shown in Fig. 1, the inactive compounds in terms of numbers, dominated the datasets. PaDEL was used to generate 4053 molecular descriptors and fingerprints. Molecular descriptors originate from mathematical steps that converts the chemical information of compounds into numerical values (Mauri et al., 2017). The descriptors show the mathematical representation of the compounds used for QSAR modelling.

The 1983 compounds were split into 1487 internal (validation and training) data and 496 external test data. For the 1487 training data, there were 456 actives and 1031 inactive. Implementation of the SMOTE scaled the active compounds to balance with the inactive data. Minority oversampling is a technique employed in predictive models where imbalance within datasets, or instances where there is low active to inactive ratio, can be rectified by adjusting the data for the active to have equal or similar quantity with the inactive (Koivu et al., 2020). With regards to the feature space, the application of variance filter (Bommert et al., 2020) for the descriptors reduced the dimensionality from 1445 to 800 features. A variance threshold of 0.1 was set to filter out descriptors with no significant difference within the data. Recursive feature elimination (RFE) (Qi et al., 2018; Escanilla et al., 2018) was used to rank the descriptors and features which were least ranked were removed. RFE wraps machine learning algorithms and it is used as the core for selecting desired features (Darst et al., 2018). Starting with all features, the core algorithm is fitted and the descriptors are ranked according to how important they affect the outcome, the least important features are removed and the model is refitted with the remaining descriptors. This process is repeated until desired number of descriptors remains. A total of 612 descriptors were used in developing the predictive models with the top ranked 20 shown in Fig. 2. These features included atomic count, atom type electro topological state, bond count, Crippen logP, molecular distance edge and XLogP. The descriptors were compared to molecular properties of known Ebola antiviral compounds. The profound properties were LogP, polar surface area, number of atoms, hydrogen bond, hydrophobic features (Ekins et al., 2014), molecular weight, oxygen and nitrogen atoms, number of OH and NH_n groups, rule of 5 violation, rotatable bonds and volume (Angstroms^3) (Bartzatt, 2016). All these features were part of the computed descriptors used in developing the algorithms with number of atoms, LogP and number of bonds being part of the top ranked 20 features.

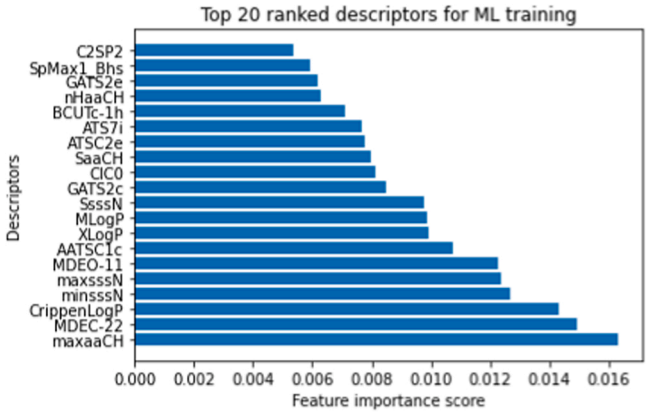


Fig. 2. A bar plot of the top ranked 20 descriptors. Overall, a total of 612 descriptors were used to implement the machine learning models.

Table 1
Results of the developed machine learning models showing the average accuracy, precision, recall, F1 scores and balanced accuracy values after cross-validation (CV) and prediction on internal test data.

Model		Accuracy	Precision	Recall	F1_Score	Balanced accuracy
KNN	CV	0.80	0.77	0.89	0.83	0.80
	Test	0.80	0.75	0.88	0.81	0.78
NB	CV	0.65	0.61	0.92	0.73	0.66
	Test	0.62	0.58	0.89	0.70	0.64
SVM	CV	0.86	0.86	0.87	0.85	0.83
	Test	0.81	0.81	0.84	0.82	0.86
RF	CV	0.89	0.88	0.91	0.90	0.89
	Test	0.86	0.82	0.84	0.85	0.87
LR	CV	0.84	0.83	0.88	0.85	0.84
	Test	0.82	0.81	0.84	0.83	0.82

3.2. Model Development and Validation

We employed 5 machine learning algorithms namely kNN, NB, SVM, RF and LR to build robust predictive models with plausible performance on the dataset for the study. The k-NN stores training data to make predictions on a query by classifying the data point based on the majority class label of its k number of nearest neighbors (Zhang et al., 2018). The GaussianNB classifier is centered on the Bayes theorem. It is a form of Naïve Bayes that uses the Gaussian normal distribution (Asafu-Adjei and Betensky, 2015). SVM is one of the simplest forms of supervised machine learning algorithms used to classify data labels by generating a hyperplane that best separate the data labels (Kramer, 2016). Random forest classifier consists of an ensemble of decision trees (Paul et al., 2018). Prediction is based on aggregation of votes from individual trees. Logistic regression is a statistical classification algorithm that predicts the probability of a dependent variable by examining

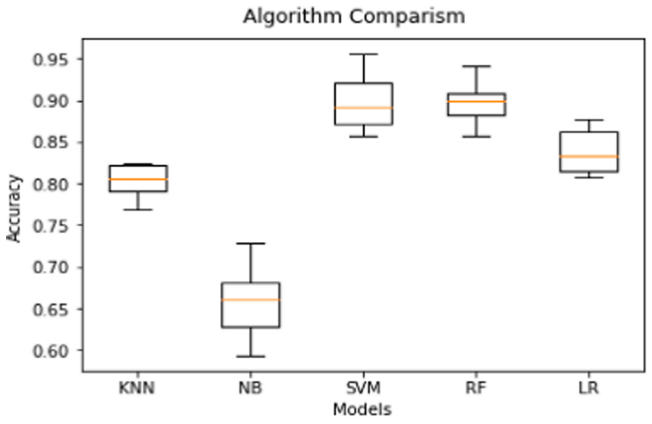


Fig. 3. A boxplot of average accuracy from the 10-fold cross-validation.

the relationships within the independent variables (Stoltzfus, 2011).

Random Forest model produced the best results from the cross-validation (Table 1 and Fig. 3). For each of the classification metrics other than recall, random forest model outperformed the rest of the predictive models. Same trends in results were observed when the predictive models were tested on the internal test data. The logistic regression model followed keenly in performance to the RF model with accuracy, precision, and recall scores of 0.82, 0.81 and 0.84, respectively from the cross-validation.

Since numerous antiviral ML predictive models have been developed using similar classifiers (Gawriljuk et al., 2021; Janairo et al., 2021; Choi et al., 2021; Gupta and Mohanty, 2021; Ekins et al., 2015; Rajput and Kumar, 2021; Sandhu et al., 2021), we compared their performance to our developed models, though none of them was trained on the cell entry inhibitor datasets used for this study. The area under the curve (AUC) of the receiver operating characteristic curve (ROC curve), accuracy, and F1 score were 0.95 (Fig. 6), 0.89 and 0.9 (Table 1), respectively. This performance is similar or better compared to others (Sandhu et al., 2021; Agyapong et al., 2021). A Bayesian classifier (Ekins et al., 2015) was trained on 868 compounds from EBOV replication and pseudotype entry assay to identify active compounds against Ebola virus, the Bayesian model performed well with an ROC score of 0.86. A support vector machine model (Sandhu et al., 2021) built to classify inhibitors trained on 5692 molecules from BindingDB database

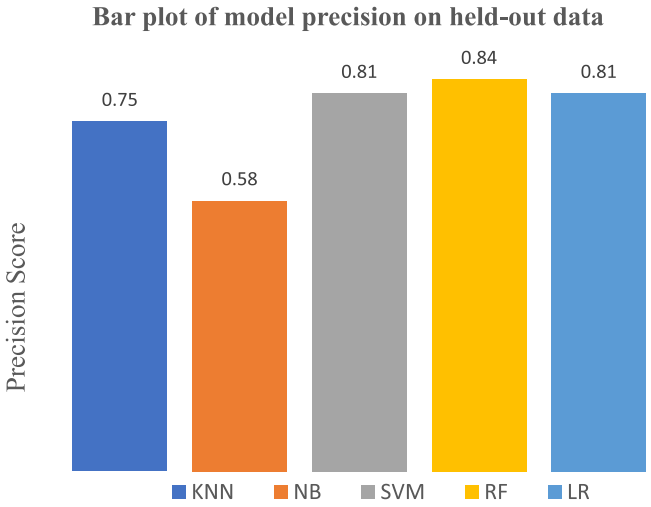


Fig. 5. Bar plot of model precision on the external datasets. Random forest model was the most precise followed by support vector machine model and the logistic regression.

produced an overall accuracy of 85.38 %. A Proteochemometric-based SVM (Agyapong et al., 2021) trained with bioactive compounds from BindingDB database yielded a performance accuracy of 93 % and ROC-AUC score of 87 %. Other than the aforementioned, regression-based predictive models comprising support vector machine, random forest and artificial neural network accessible via a web server, computes inhibition efficiencies of anti-EBOV compounds with Pearson's correlation coefficient ranging from 0.83 to 0.98 (Rajput and Kumar, 2021). These models were developed from 305 anti-EBOV compounds from DrugRepV (Rajput et al., 2021) using the respective IC_{50} values. Predictions made by these regressions based anti-Ebola webserver is a continuous variable of IC_{50} values of query compounds as compared to our developed model which gives a discrete set of predictions, distinguishing compounds as either active or inactive to blocking the entry of Ebola virus to the host cell. The ROC curves for the rest of the predictive models have been generated (Supplementary Figs. S1 to S4). The ROC curve is generated by plotting true positive rate (TPR) against false positive rate (FPR) (Fawcett, 2006). The curve summarizes the confusion matrix that each threshold produces. The

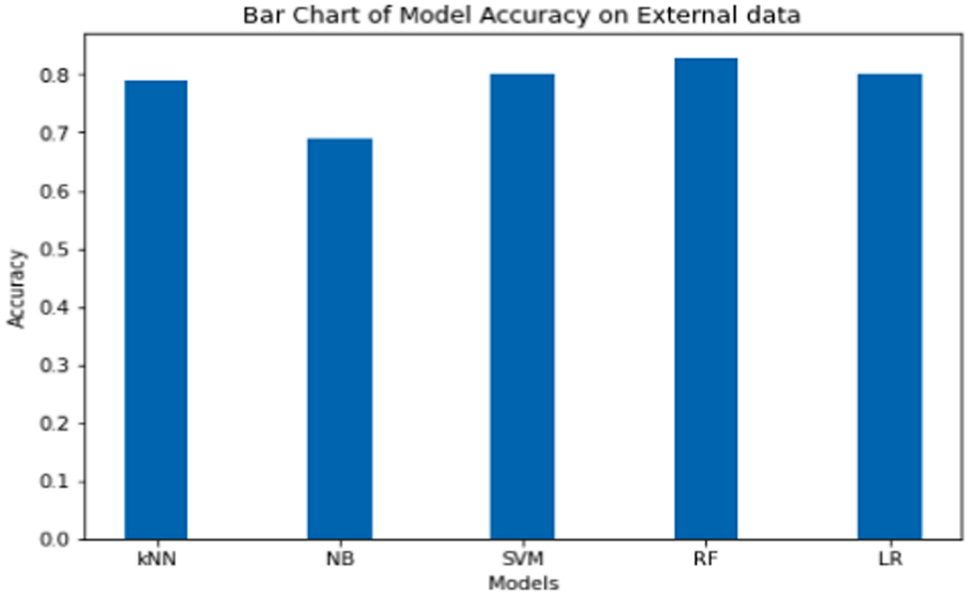


Fig. 4. A bar plot showing model accuracy on external data.

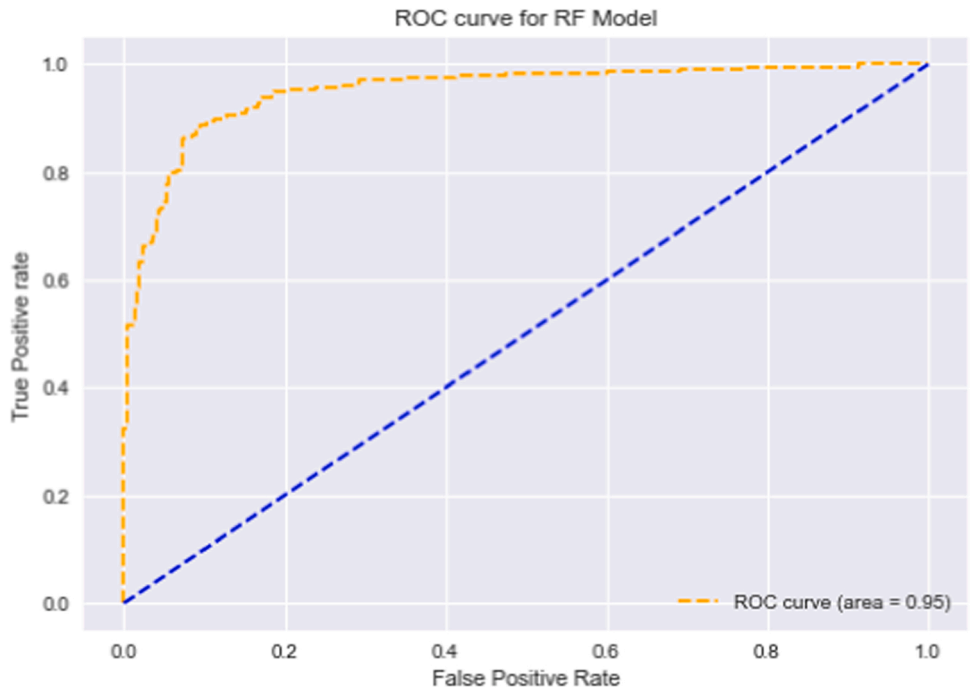


Fig. 6. Receiver operating characteristic curve (ROC curve) of true positive rate (TPR) against false positive rate (FPR). TPR indicates the ratio of active compounds that were correctly predicted to all positive observations. FPR represents the fraction of inactive compounds that were predicted as active to all negative observations.

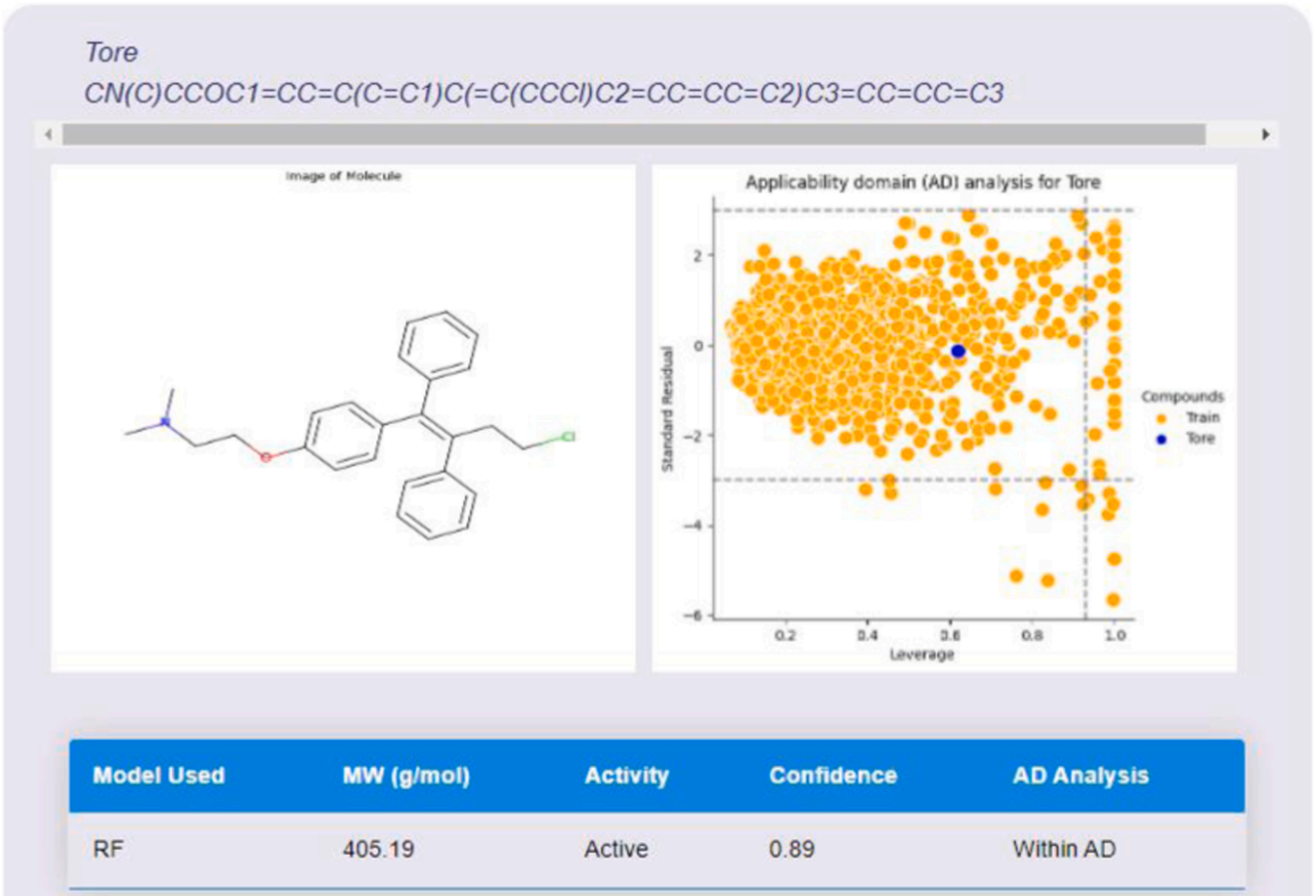


Fig. 7. A web interface of the EBOLApred. The results of the query of toremifene predicted as active with confidence score of 0.89 and molecular weight (MW) of 405.19 g/mol using random forest model.

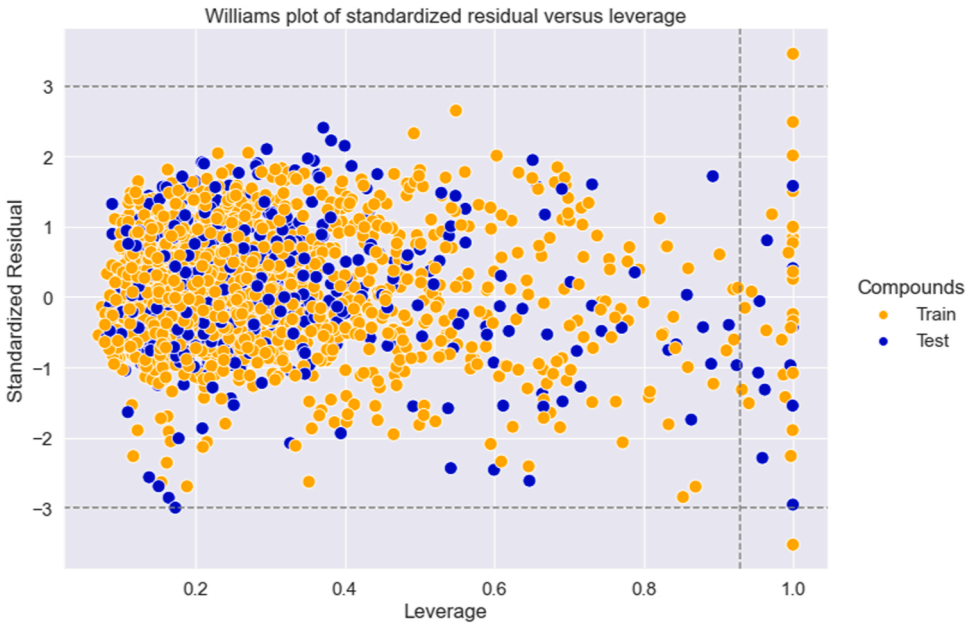


Fig. 8. A plot of standardized residual versus leverage showing the boundaries defining the domain of applicability of the predictive models. The horizontal grey line represents the residual threshold value of ± 3. The leverage threshold was computed to be 0.93 and it is shown as the vertical grey line. In total, 37 compounds consisting of 26 training and 11 test samples were outliers.

closer the ROC curve to the diagonal, where TPR = FPR, the less accurate the predictive model. From the Fig. 6, the ROC curve is to the left of the diagonal. This shows that the proportion of correctly classified active compounds (TPR) is greater than the proportion of compounds that were incorrectly classified as inactive (FPR). An AUC score ranges from 0 to 1 showing the probability of a model to rank a randomly chosen active compound higher than a randomly chosen inactive compound. When the AUC is close to 1, the model has a reasonable ability to classify active from inactive compounds (Fawcett, 2006).

3.3. Prediction on held-out data

An external data comprising of 496 compounds were held out and remained untouched throughout the training and development of the predictive models. This data provides another platform in evaluating the developed models. The performance metrics of the models were again analyzed using the held out external test data. Despite a drop in accuracy on the test data (Fig. 4) in comparison to the results from the cross-validation (Fig. 3), the robustness of the random forest model was reaffirmed. The RF model was more accurate than the rest of the predictive models with an accuracy score of 0.83. Fig. 5 illustrate the models' precisions on the held-out data. The random forest model was again, more precise to the rest of the models. With all performance metrics, the Naïve Bayes model was the least.

3.4. Application of the model

To further validate and apply the predictive models on experimentally validated EBOV VP40 inhibitors, the random forest model was used to predict the bioactivity of two query molecules, tilorone and toremifene which had experimentally been shown to inhibit the matrix protein with IC_{50} values of 3.43 and 0.56 μM, respectively (Kouznetsova et al., 2014). The random forest model predicted tilorone and toremifene as active with confidence scores of 0.75 and 0.89 (Fig. 7), respectively.

3.5. Applicability domain analysis

The performance of machine learning-based quantitative structure activity relationship (QSAR) models on a query is dependent on the similarity of the query molecule to the training examples (Rakhimbekova et al., 2020). Applicability Domain (AD) of a machine learning model defines an area of chemical space within the training data for which the developed model gives steady and reliable predictions (Roy et al., 2015). A Williams plot (Kar et al., 2018) of standardized residual and leverage values were constructed to determine the domain of applicability of the QSAR models. The residual and leverage values were computed from the 612 features used to build the predictive models. The AD was defined using a leverage threshold and standardized residual boundaries (Fig. 8).

Leverage threshold value (h) was determined by (Kar et al., 2018);

$$h = \frac{3(p+1)}{n} \tag{8}$$

where p is the number of descriptors used and n is the number of training compounds.

3.6. Model deployment as a web application

Random forest, logistic regression and support vector machine models were integrated into a webserver to enable user prediction of potential anti-EBOV molecules. The server known as EBOLApred is freely available at http://197.255.126.13:8000/. By querying a compound in a SMILES format, it predicts whether the compound is an EBOV cell entry inhibitor.

3.7. Application of models in anti-EBOV drug repurposing

Computational approaches have surged in recent years to unveil novel biotherapeutic compounds to combat EVD. Molecular docking studies, molecular dynamic simulations, Bayesian models and anti-Ebola algorithms have thus been exploited (Kwofie et al., 2019b; Ekins et al., 2015; Rajput and Kumar, 2021). Preventing EBOV from infecting host cells is a promising antiviral approach. As such, in the current study, we developed machine learning-based models to identify potential anti-EBOV agents. Three of the five constructed ML classifiers comprising random forest, support vector machine and logistic

regression with excellent accuracy, precision and recall scores were implemented as a webserver. EBOLApred and its accompanying algorithms allow the large-scale computational screening of compound libraries to prioritize potential anti-EBOV molecules or plausible cell entry inhibitors for experimental characterization. Food and Drug Agency (FDA) approved drugs (Lane et al., 2020; Muthaiyan et al., 2021) can be screened with the algorithms to identify potential anti-EBOV candidates for pharmacological evaluation.

4. Conclusion

The study developed five machine learning algorithms comprising k-nearest neighbor, gaussian naïve bayes, support vector machines, random forest and logistic regression using bioactive datasets from PubChem database to predict EBOV cell entry inhibitors. The 1d and 2d descriptors as well as fingerprints of the compounds were calculated and after feature selection through variance filter and recursive feature elimination, 612 out of 4053 descriptors and fingerprints were used in training the predictive models. Random forest emerged as the best model with an accuracy of 0.89 and ROC-AUC score of 0.95. The RF model together with LR and SVM models were deployed as a web application with an enhanced user query interface to support the prediction of EBOV cell entry inhibitors. The robust and efficient models developed augment the prediction of anti-EBOV compounds with chemotherapeutic potential.

CRediT authorship contribution statement

S.K.K, J.A and M.D.W conceptualized and designed the project. J.A and S.K.K developed the models with input from O.A, K.A. and M.D. J.A wrote the first draft of the manuscript with inputs from the others. All the authors approved the submission of the final draft.

Funding

This research received no external funding.

Data availability

All data IDs for this work are provided in the manuscript and the models are available at https://github.com/ebolapred/EBOV.git.

Conflicts of interest

The authors declare no conflict of interest.

Acknowledgments

The authors are grateful to the University of Ghana Computing Systems for hosting the web server. We are also thankful to the West African Centre for Cell Biology of Infectious Pathogens for providing the Zuputo computing infrastructure for the work.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.compbiolchem.2022.107766.

References

Agyapong, O., Miller, W.A., Wilson, M.D., Kwofie, S.K., 2021. Development of a proteochemometric-based support vector machine model for predicting bioactive molecules of tubulin receptors. Mol. Divers. https://doi.org/10.1007/s11030-021-10329-w.

Ahsan, M.M., Mahmud, M.A.P., Saha, P.K., Gupta, K.D., Siddique, Z., 2021. Effect of data scaling methods on machine learning algorithms and model performance. Technologies 9 (3), 52. https://doi.org/10.3390/technologies9030052.

Alam El-Din, H.M., et al., 2016. Molecular docking based screening of compounds against VP40 from Ebola virus. Bioinformation 12 (3), 192–196. https://doi.org/10.6026/97320630012192.

Asafu-Adjei, J.K., Betensky, R.A., 2015. A Pairwise Naïve Bayes approach to Bayesian classification. Int. J. Pattern Recognit. Artif. Intell. 29 (7) https://doi.org/10.1142/S0218001415500238.

Asiedu, S.O., Kwofie, S.K., Broni, E., Wilson, M.D., 2021. Computational identification of potential anti-inflammatory natural compounds targeting the p38 Mitogen-Activated Protein Kinase (MAPK): implications for COVID-19-induced cytokine storm. Biomolecules 11 (5). https://doi.org/10.3390/biom11050653.

Bartzatt, R., 2016. "Properties and drug-likeness of compounds that inhibit Ebola Virus Disease (EVD. Int. J. Trop. Dis. Heal. 15 (2), 1–17. https://doi.org/10.9734/ijtdh/2016/25021.

Barua, S., Islam, M.M., Murase, K., 2011. A novel synthetic minority oversampling technique for imbalanced data set learning. Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.) 7063 LNCS (2), 735–744. https://doi.org/10.1007/978-3-642-24958-7_85.

D. Berrar, "Cross-validation," Encycl. Bioinforma. Comput. Biol. ABC Bioinforma., vol. 1–3, no. January 2018, pp. 542–545, 2018, doi: 10.1016/B978–0-12–809633-8.20349-X.

Bhagwati, S., Siddiqi, M.I., 2020. Deep neural network modeling based virtual screening and prediction of potential inhibitors for renin protein. J. Biomol. Struct. Dyn. 1–14. https://doi.org/10.1080/07391102.2020.1860825.

Bommert, A., Sun, X., Bischl, B., Rahnenführer, J., Lang, M., 2020. Benchmark for filter methods for feature selection in high-dimensional classification data. Comput. Stat. Data Anal. 143, 106839 https://doi.org/10.1016/j.csda.2019.106839.

Bornholdt, Z.A., et al., 2013. XStructural rearrangement of ebola virus vp40 begets multiple functions in the virus life cycle. Cell 154 (4). https://doi.org/10.1016/j.cell.2013.07.015.

Capuzzi, S.J., et al., 2018. Computer-aided discovery and characterization of novel Ebola virus inhibitors. J. Med. Chem. 61 (8), 3582–3594. https://doi.org/10.1021/acs.jmedchem.8b00035.

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2011. SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. 16, 321–357. https://doi.org/10.1613/jair.953.

Chen, Q., Meng, Z., Liu, X., Jin, Q., Su, R., 2018. Decision variants for the automatic determination of optimal feature subset in RF-RFE. Genes (Basel) 9 (6), 10.3390/genes9060301.

Choi, J., et al., 2021. Prediction of African swine fever virus inhibitors by molecular docking-driven machine learning models. Molecules 26 (12). https://doi.org/10.3390/molecules26123592.

Darko, L.K.S., Broni, E., Amuzu, D.S.Y., Wilson, M.D., Parry, C.S., Kwofie, S.K., 2021. Computational study on potential novel anti-Ebola virus protein VP35 natural compounds. Biomedicines 9 (12). https://doi.org/10.3390/biomedicines9121796.

Darst, B.F., Malecki, K.C., Engelman, C.D., 2018. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. BMC Genet 19 (Suppl. 1), 1–6. https://doi.org/10.1186/s12863-018-0633-8.

Davey, R.T.J., et al., 2016. A randomized, controlled trial of ZMapp for Ebola virus infection. N. Engl. J. Med. 375 (15), 1448–1456. https://doi.org/10.1056/NEJMoa1604330.

Dhama, K., et al., 2018. Advances in designing and developing vaccines, drugs, and therapies to counter Ebola virus. Front. Immunol. 9, 1803. https://doi.org/10.3389/fimmu.2018.01803.

Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M., 2006. Review: a gentle introduction to imputation of missing values. J. Clin. Epidemiol. 59 (10), 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014.

Ekins, S., Freundlich, J.S., Coffee, M., 2014. A common feature pharmacophore for FDA-approved drugs inhibiting the Ebola virus. F1000Research 3, 277. https://doi.org/10.12688/f1000research.5741.2.

Ekins, S., Freundlich, J.S., Clark, A.M., Anantpadma, M., Davey, R.A., Madrid, P., 2015. Machine learning models identify molecules active against the Ebola virus in vitro. F1000Research 4, 1091. https://doi.org/10.12688/f1000research.7217.3.

Elreedy, D., Atiya, A.F., 2019. A comprehensive analysis of synthetic minority oversampling technique (SMOTE) for handling class imbalance. Inf. Sci. (N.Y.). 505, 32–64. https://doi.org/10.1016/j.ins.2019.07.070.

J. Emanuel, A. Marzi, and H. Feldmann, Chapter Nine - Filoviruses: Ecology, Molecular Biology, and Evolution, vol. 100, M. Kielian, T. C. Mettenleiter, and M. J. B. T.-A. in V. R. Roossinck, Eds. Academic Press, 2018, pp. 189–221. doi: https://doi.org/10.1016/bs.aivir.2017.12.002.

N.S. Escanilla, L. Hellerstein, R. Kleiman, Z. Kuang, J.D. Shull, and D. Page, "Recursive Feature Elimination by Sensitivity Testing.," Proc. Int. Conf. Mach. Learn. Appl. Int. Conf. Mach. Learn. Appl., vol. 2018, pp. 40–47, Dec. 2018, doi: 10.1109/ICMLA.2018.00014.

Espinoza, G.Z., Angelo, R.M., Oliveira, P.R., Honorio, K.M., 2021. Evaluating Deep Learning models for predicting ALK-5 inhibition. PLoS One 16 (1), e0246126. https://doi.org/10.1371/journal.pone.0246126.

Fabianpedregosa, F.P., et al., 2011. Scikit-learn: machine learning in Python Gaël varoquaux bertrand thirion vincent Dubourg Alexandre Passos PEDREGOSA, VAROQUAUX, GRAMFORT ET AL. Matthieu Perrot. J. Mach. Learn. Res. 12, 2825–2830. https://doi.org/10.5555/1953048.

Fawcett, T., 2006. An introduction to ROC analysis. Pattern Recognit. Lett. 27 (8), 861–874. https://doi.org/10.1016/j.patrec.2005.10.010.

García, V., Mollineda, R.A., Sánchez, J.S., 2009. Index of balanced accuracy: a performance measure for skewed class distributions. Lect. Notes Comput. Sci. (Incl. Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinforma.) 5524 LNCS, 441–448. https://doi.org/10.1007/978-3-642-02172-5_57.

Gawriljuk, V.O., et al., 2021. Machine learning models identify inhibitors of SARS-CoV-2. J. Chem. Inf. Model. 61 (9), 4224–4235. https://doi.org/10.1021/acs.jcim.1c00683.
Gupta, P., Mohanty, D., 2021. SMMPPI: a machine learning-based approach for prediction of modulators of protein-protein interactions and its application for identification of novel inhibitors for RBD:hACE2 interactions in SARS-CoV-2. Brief. Bioinform. 22 (5) https://doi.org/10.1093/bib/bbab111.
Haneczok, J., Delijewski, M., 2021. Machine learning enabled identification of potential SARS-CoV-2 3CLpro inhibitors based on fixed molecular fingerprints and Graph-CNN neural representations. J. Biomed. Inform. 119, 103821 https://doi.org/10.1016/j.jbi.2021.103821.
Hansen, F., Feldmann, H., Jarvis, M.A., 2021. Targeting Ebola virus replication through pharmaceutical intervention. Expert Opin. Investig. Drugs 30 (3), 201–226. https://doi.org/10.1080/13543784.2021.1881061.
Jacob, S.T., et al., 2020. Ebola virus disease. Nat. Rev. Dis. Prim. 6 (1), 13. https://doi.org/10.1038/s41572-020-0147-3.
Jain, S., Martynova, E., Rizvanov, A., Khaiboullina, S., Baranwal, M., 2021. Structural and functional aspects of ebola virus proteins. Pathogens 10 (10), 1–29. https://doi.org/10.3390/pathogens10101330.
Janairo, G.I.B., Yu, D.E.C., Janairo, J.I.B., 2021. A machine learning regression model for the screening and design of potential SARS-CoV-2 protease inhibitors. Netw. Model. Anal. Heal. Inform. Bioinforma. 10 (1), 51. https://doi.org/10.1007/s13721-021-00326-2.
Jerez, J.M., et al., 2010. Missing data imputation using statistical and machine learning methods in a real breast cancer problem. Artif. Intell. Med. 50 (2), 105–115. https://doi.org/10.1016/j.artmed.2010.05.002.
Jiao, Y., Du, P., 2016. Performance measures in evaluating machine learning based bioinformatics predictors for classifications. Quant. Biol. 4 (4), 320–330. https://doi.org/10.1007/s40484-016-0081-2.
Johnson, R.F., McCarthy, S.E., Godlewski, P.J., Harty, R.N., 2006. Ebola virus VP35-VP40 interaction is sufficient for packaging ^{3}E–^{5}E minigenome RNA into virus-like particles. J. Virol. 80 (11), 5135–5144. https://doi.org/10.1128/jvi.01857-05.
Juneau, J., Baker, J., Ng, V., Soto, L., Wierzbicki, F., 2010. Web Applications With Django. https://doi.org/10.1007/978-1-4302-2528-7_14.
Kar, S., Roy, K., Leszczynski, J., 2018. Applicability domain: a step toward confident predictions and decidability for QSAR modeling. Methods Mol. Biol. 1800, 141–169. https://doi.org/10.1007/978-1-4939-7899-1_6.
Khan, S., Fakhar, Z., Ahmad, A., 2021. Targeting Ebola virus VP40 protein through novel inhibitors: exploring the structural and dynamic perspectives on molecular landscapes. J. Mol. Model. 27 (2), 49. https://doi.org/10.1007/s00894-021-04682-8.
Koch, L.K., Cunze, S., Kochmann, J., Klimpel, S., 2020. Bats as putative Zaire ebolavirus reservoir hosts and their habitat suitability in Africa. Sci. Rep. 10 (1), 14268. https://doi.org/10.1038/s41598-020-71226-0.
Koivu, A., Sairanen, M., Airola, A., Pahikkala, T., 2020. Synthetic minority oversampling of vital statistics data with generative adversarial networks. J. Am. Med. Inform. Assoc. 27 (11), 1667–1674. https://doi.org/10.1093/jamia/ocaa127.
Kouznetsova, J., et al., 2014. Identification of 53 compounds that block Ebola virus-like particle entry via a repurposing screen of approved drugs. Emerg. Microbes \ Infect. 3 (1), 1–7. https://doi.org/10.1038/emi.2014.88.
Kramer, O., 2016. Machine learning for evolution strategies.
Krishnasamy, L., Saikumar, C., 2015. Updates on treatment of ebola virus disease. Malays. J. Med. Sci. 22 (6), 54–57.
J. Kuhn et al., Filoviridae, 2020.
Kumar, P., Bhatnagar, R., Gaur, K., Bhatnagar, A., 2021. Classification of imbalanced data: review of methods and applications. IOP Conf. Ser. Mater. Sci. Eng. 1099 (1), 012077 https://doi.org/10.1088/1757-899x/1099/1/012077.
Kumari, M., Subbarao, N., 2021. Deep learning model for virtual screening of novel 3C-like protease enzyme inhibitors against SARS coronavirus diseases. Comput. Biol. Med. 132, 104317 https://doi.org/10.1016/j.compbiomed.2021.104317.
Kwofie, S.K., et al., 2019a. Pharmacoinformatics-based identification of potential bioactive compounds against Ebola virus protein VP24. Comput. Biol. Med. 113, 103414 https://doi.org/10.1016/j.compbiomed.2019.103414.
Kwofie, S.K., et al., 2019b. Pharmacoinformatics-based identification of potential bioactive compounds against Ebola virus protein VP24. Comput. Biol. Med. vol. 113 (August) https://doi.org/10.1016/j.compbiomed.2019.103414.
Kwofie, S.K., et al., 2021. Cheminformatics-Based Identification of Potential Novel Anti-SARS-CoV-2 Natural Compounds of African Origin. Molecules vol. 26 (2). https://doi.org/10.3390/molecules26020406.
Lane, T.R., et al., 2020. Repurposing Pyramax®, quinacrine and tilorone as treatments for Ebola virus disease. Antivir. Res 182. https://doi.org/10.1016/j.antiviral.2020.104908.
Lane, T.R., Ekins, S., 2020. Toward the target: tilorone, quinacrine, and pyronaridine bind to Ebola virus glycoprotein. ACS Med. Chem. Lett. 11 (8), 1653–1658. https://doi.org/10.1021/acsmedchemlett.0c00298.
Lee, A., 2021. Ansuvimab: first approval. Drugs 81 (5), 595–598. https://doi.org/10.1007/s40265-021-01483-4.
Lee, J.E., Fusco, M.L., Hessell, A.J., Oswald, W.B., Burton, D.R., Saphire, E.O., 2008. Structure of the Ebola virus glycoprotein bound to an antibody from a human survivor. Nature 454 (7201), 177–182. https://doi.org/10.1038/nature07082.
Leligdowicz, A., et al., 2016. Ebola virus disease and critical illness. Crit. Care 20 (1), 217. https://doi.org/10.1186/s13054-016-1325-2.
LemaîtreLemaître, G., Nogueira, F., char, C.K.A., 2017. Imbalanced-learn: a python toolbox to tackle the curse of imbalanced datasets in machine learning. J. Mach. Learn. Res. vol. 18, 1–5. https://doi.org/10.5555/3122009.
Madara, J.J., Han, Z., Ruthel, G., Freedman, B.D., Harty, R.N., 2015. The multifunctional Ebola virus VP40 matrix protein is a promising therapeutic target. Future Virol. 10 (5), 537–546, 10.2217/fvl.15.6.
Madrid, P.B., et al., 2015. Evaluation of Ebola virus inhibitors for drug repurposing. ACS Infect. Dis. 1 (7), 317–326. https://doi.org/10.1021/acsinfecdis.5b00030.
Malik, A.A., Chotpatiwetchkul, W., Phanus-Umporn, C., Nantasenamat, C., Charoenkwan, P., Shoombuatong, W., 2021. "StackHCV: a web-based integrative machine-learning framework for large-scale identification of hepatitis C virus NS5B inhibitors. J. Comput. Aided Mol. Des. 35 (10), 1037–1053. https://doi.org/10.1007/s10822-021-00418-1.
Mauri, A., Consonni, V., Todeschini, R., 2017. Molecular descriptors. Handb. Comput. Chem. 2065–2093. https://doi.org/10.1007/978-3-319-27282-5_51.
Mekni, N., Coronnello, C., Langer, T., De Rosa, M., Perricone, U., 2021. Support vector machine as a supervised learning for the prioritization of novel potential SARS-CoV-2 main protease inhibitors. Int. J. Mol. Sci. 22 (14) https://doi.org/10.3390/ijms22147714.
Muthaiyan, M., Naorem, L.D., Seenappa, V., Pushan, S.S., Venkatesan, A., . 2021. Ebolabase: Zaire ebolavirus-human protein interaction database for drug-repurposing. Int. J. Biol. Macromol. vol. 182, 1384–1391. https://doi.org/10.1016/j.ijbiomac.2021.04.184.
Nagarajan, N., Yapp, E.K.Y., Le, N.Q.K., Yeh, H.-Y., 2019. In silico screening of sugar alcohol compounds to inhibit viral matrix protein VP40 of Ebola virus. Mol. Biol. Rep. 46 (3), 3315–3324. https://doi.org/10.1007/s11033-019-04792-w.
Odhar, H.A., Rayshan, A.M., Ahjel, S.W., Hashim, A.A., Albeer, A.A.M.A., 2019. Molecular docking enabled updated screening of the matrix protein VP40 from Ebola virus with millions of compounds in the MCULE database for potential inhibitors. Bioinformation 15 (9), 627–632. https://doi.org/10.6026/97320630015627.
Osterholm, M.T., et al., 2015. Transmission of Ebola viruses: what we know and what we do not know. MBio 6 (2). https://doi.org/10.1128/mBio.00137-15.
Paul, A., Mukherjee, D.P., Das, P., Gangopadhyay, A., Chintha, A.R., Kundu, S., 2018. Improved random forest for classification. IEEE Trans. Image Process. Publ. IEEE Signal Process. Soc. 27 (8), 4012–4024. https://doi.org/10.1109/TIP.2018.2834830.
Perez-Ortiz, M., Gutierrez, P.A., Tino, P., Hervas-Martinez, C., 2016. "Oversampling the minority class in the feature space. IEEE Trans. Neural Netw. Learn. Syst. 27 (9), 1947–1961. https://doi.org/10.1109/TNNLS.2015.2461436.
Qi, C., Meng, Z., Liu, X., Jin, Q., Su, R., 2018. Decision variants for the automatic determination of optimal feature subset in RF-RFE. Genes (Basel) 9, 301. https://doi.org/10.3390/genes9060301.
A.I. Qureshi, Chapter 3 - Ebola Virus: The Origins, A. I. B. T.-E. V. D. Qureshi, Ed. Academic Press, 2016, pp. 23–37. doi: https://doi.org/10.1016/B978–0-12–804230-4.00003–0.
Rajak, H., Jain, D.K., Singh, A., Sharma, A.K., Dixit, A., 2015. Ebola virus disease: past, present and future. Asian Pac. J. Trop. Biomed. 5 (5), 337–343 https://doi.org/10.1016/S2221-1691(15)30365-8.
Rajput, A., Kumar, M., 2021. Anti-Ebola: an initiative to predict Ebola virus inhibitors through machine learning. Mol. Divers. 1–10. https://doi.org/10.1007/s11030-021-10291-7.
Rajput, A., Kumar, A., Megha, K., Thakur, A., Kumar, M., 2021. DrugRepV: a compendium of repurposed drugs and chemicals targeting epidemic and pandemic viruses. Brief. Bioinform. 22 (2), 1076–1084. https://doi.org/10.1093/bib/bbaa421.
Rakhimbekova, A., Madzhidov, T.I., Nugmanov, R.I., Gimadiev, T.R., Baskin, I.I., Varnek, A., 2020. Comprehensive analysis of applicability domains of QSPR models for chemical reactions. Int. J. Mol. Sci. 21 (15), 1–20. https://doi.org/10.3390/ijms21155542.
Roy, K., Kar, S., Ambure, P., 2015. On a simple approach for determining applicability domain of QSAR models. Chemom. Intell. Lab. Syst. 145, 22–29. https://doi.org/10.1016/j.chemolab.2015.04.013.
Salata, C., Calistri, A., Alvisi, G., Celestino, M., Parolin, C., Palù, G., 2019. Ebola virus entry: from molecular characterization to drug discovery. Viruses 11 (3). https://doi.org/10.3390/v11030274.
Sandhu, H., Kumar, R.N., Garg, P., 2021. Machine learning-based modeling to predict inhibitors of acetylcholinesterase. Mol. Divers. (0123456789) https://doi.org/10.1007/s11030-021-10223-5.
Sankar, M., K, L., Jeyachandran, S., Pandi, B., 2021. Screening of inhibitors as potential remedial against Ebolavirus infection: pharmacophore-based approach. J. Biomol. Struct. Dyn. 39 (2), 395–408. https://doi.org/10.1080/07391102.2020.1715260.
Schuler, J., Hudson, M.L., Schwartz, D., Samudrala, R., 2017. A systematic review of computational drug discovery, development, and repurposing for Ebola virus disease treatment. Molecules 22 (10). https://doi.org/10.3390/molecules22101777.
Seliya, N., Khoshgoftaar, T.M., Van Hulse, J., 2009. A study on the relationships of classifier performance metrics. Proc. - Int. Conf. Tools Artif. Intell. ICTAI 59–66. https://doi.org/10.1109/ICTAI.2009.25.
Shockley, K.R., 2016. Estimating potency in high-throughput screening experiments by maximizing the rate of change in weighted Shannon entropy. Sci. Rep. vol. 6 (1), 27897. https://doi.org/10.1038/srep27897.
Sissoko, D., et al., 2016. Experimental treatment with favipiravir for Ebola virus disease (the JIKI Trial): a historically controlled, single-arm proof-of-concept trial in Guinea. PLoS Med 13 (3), e1001967. https://doi.org/10.1371/journal.pmed.1001967.
Sliwoski, G., Kothiwale, S., Meiler, J., Lowe Jr, E.W., 2013. Computational methods in drug discovery. Pharmacol. Rev. 66 (1), 334–395. https://doi.org/10.1124/pr.112.007336.
J.C. Stoltzfus, Logistic regression: a brief primer., Acad. Emerg. Med. Off. J. Soc. Acad. Emerg. Med., vol. 18, no. 10, pp. 1099–1104, Oct. 2011, doi: 10.1111/j.1553–2712.2011.01185.x.

Tamilvanan, T., Hopper, W., 2013. High-throughput virtual screening and docking studies of matrix protein vp40 of ebola virus. Bioinformation 9 (6), 286–292. https://doi.org/10.6026/97320630009286.

Taneja, S., Gupta, P.R., 2014. Python as a tool for web server application development. Int. J. Inf., Commun. Comput. Technol. 2 (1), 77–83. ⟨https://www.jimsindia.org/8i_Journal/VolumeII/Python-as-a-tool-for-web-server-application-development.pdf⟩.

Tendong, W., Lebrun, P., Verbist, B., 2020. Controlling the reproducibility of AC50 estimation during compound profiling through Bayesian β-expectation tolerance intervals. SLAS Disco 25 (9), 1009–1017. https://doi.org/10.1177/2472555220918201.

Tompa, D.R., Immanuel, A., Srikanth, S., Kadhirvel, S., 2021. Trends and strategies to combat viral infections: a review on FDA approved antiviral drugs. Int. J. Biol. Macromol. 172, 524–541. https://doi.org/10.1016/j.ijbiomac.2021.01.076.

Tougui, I., Jilbab, A., El Mhamdi, J., 2021. Impact of the choice of cross-validation techniques on the results of machine learning-based diagnostic applications. Healthc. Inform. Res. 27 (3), 189–199. https://doi.org/10.4258/hir.2021.27.3.189.

Toussi, C.A., Haddadnia, J., Matta, C.F., 2021. Drug design by machine-trained elastic networks: predicting Ser/Thr-protein kinase inhibitors' activities. Mol. Divers. 25 (2), 899–909. https://doi.org/10.1007/s11030-020-10074-6.

S. bin Uzayr, N. Cloud, and T. Ambler, "React BT - JavaScript Frameworks for Modern Web Development: The Essential Frameworks, Libraries, and Tools to Learn Right Now," pp. 507–521, 2019, [Online]. Available: https://doi.org/10.1007/978–1-4842–4995-6_13.

Velliangiri, S., Alagumuthukrishnan, S., Thankumar joseph, S.I., 2019. A review of dimensionality reduction techniques for efficient computation. Procedia Comput. Sci. 165, 104–111. https://doi.org/10.1016/j.procs.2020.01.079.

Virtanen, P., et al., 2020. SciPy 1.0: fundamental algorithms for scientific computing in Python. Nat. Methods 17 (3), 261–272. https://doi.org/10.1038/S41592-019-0686-2.

Wan, W., et al., 2017. Structure and assembly of the Ebola virus nucleocapsid. Nature 551 (7680), 394–397. https://doi.org/10.1038/nature24490.

Wang, Y., Xiao, J., Suzek, T.O., Zhang, J., Wang, J., Bryant, S.H., 2009. PubChem: a public information system for analyzing bioactivities of small molecules. Nucleic Acids Res 37 (Web Server issue), W623–W633. https://doi.org/10.1093/nar/gkp456.

Warren, T.K., et al., 2016. Therapeutic efficacy of the small molecule GS-5734 against Ebola virus in rhesus monkeys. Nature 531 (7594), 381–385. https://doi.org/10.1038/nature17180.

Wei, Q., Dunbrack Jr, R.L., 2013. The role of balanced training and testing data sets for binary classifiers in bioinformatics. PLoS One 8 (7), 1–12. https://doi.org/10.1371/journal.pone.0067863.

Yap, C.W., 2011. PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem. 32 (7), 1466–1474. https://doi.org/10.1002/JCC.21707.

Zawilińska, B., Kosz-Vnenchak, M., 2014. General introduction into the Ebola virus biology and disease. Folia Med. Cracov. 54 (3), 57–65.

Zhang, S., Li, X., Zong, M., Zhu, X., Wang, R., 2018. Efficient kNN classification with different numbers of nearest neighbors. IEEE Trans. Neural Netw. Learn. Syst. 29 (5), 1774–1785. https://doi.org/10.1109/TNNLS.2017.2673241.

Zhou, J., et al., 2021. Classification and design of HIV-1 integrase inhibitors based on machine learning. Comput. Math. Methods Med. 2021, 5559338. https://doi.org/10.1155/2021/5559338.