ONLINE ISSN :3007-309X
PRINT ISSN :3007-3081

Journal of Medical & Health Sciences Review

IDENTIFICATION OF HEMAGGLUTININ USING SEQUENCE INFORMATION MACHINE LEARNING LIGHTGBM

Rahu Sikander[1a]*, Narmeen Zakaria Bawany[1b], Ali Ghulam[2], Mujeebu Rehman[3], Rasulov Farruhbek[4]

[1a,1b]Department of Computer Science & Software Engineering Jinnah University for Women, Sindh, Pakistan, [1a]Email: rahu.sikander@juw.edu.pk
[2]Information Technology Centre, Sindh Agriculture University, Sindh, Pakistan, Email: garahu@sau.edu.pk
[3]School of Information and Communication Engineering, Guilin University of Electronic Technology, Guilin, China
[4]Department of Pharmaceutical Sciences, Andijan State Medical Institute, Andijan, Uzbekistan

ARTICLE INFO:

Keywords: Feature engineering, Sequence analysis, DDE-PSSM, LightGBM, Model accuracy

Corresponding Author:Rahu Sikander, Department of Computer Science & Software Engineering Jinnah University for Women, Sindh, Pakistan, Email: rahu.sikander@juw.edu.pk

Article History:
Published on July 2025

ABSTRACT

Introduction: HA contributes to viral infection by making it possible for the virus to fuse with the membrane of the host cell. Since HA is essential in influenza virus infection, pharmaceutical companies are focused on making drugs and vaccines against it. For this reason, it is very important to accurately identify HA for the progress of vaccination treatment. Even so, the complete identification of HA using computational methods is not enough. The purpose of this study is to build a model that helps find HA.

Methods: For this study, a benchmark dataset containing 106 HA and 106 non-HA sequences was obtained from UniProt. The samples were generated with various sequence-related properties. We developed an ensemble classifier through stacking technique by tuning features including stacking up of four machine learning (ML) methods.

Results and discussion: The accuracy of the model was found to be 97.80% in 5-fold cross-validation, whereas the area under the receiver operating characteristic (ROC) curve was 0.9930. The accuracy of the model in the test dataset was 93.18%, while the area of the ROC curve was 0.9793. Using DDE-PSSM, the LightGBM algorithm was able to achieve an accuracy of 97.13%, precision of 100.0%, sensitivity of 94.29%, specificity of 100.0%, MCC of 94.55% and an AUC of 98.30%. The accuracy, precision, sensitivity, specificity, MCC and AUC values for using anti-hypertensives were 97.80%, 100.0%, 94.10%, 100.0%, 94.51% and 99.30%, respectively. The model is presented as a particularly good predictor. The model is very useful for biochemists to explore for studying HA.

1. INTRODUCTION

Flu causes many problems in people's health and leads to varying amounts of disease worldwide [1-2]. The presence of hemagglutinin (HA), a glycoprotein on the virus surface, helps the virus enter the host cells by latching onto the cells' sialic acid [4]. It has been noted that the conserved region in hemagglutinin (HA) offers a good target for an influenza vaccine for everyone [4]. Correctly identifying the HA is important for creating helpful vaccines and therapies. Now that advanced technologies and machine learning algorithms exist, sequence-based approaches have made identifying proteins faster and simpler. It allows specific proteins to be grouped and identified using protein sequence coding methods and machine learning algorithms which are widely used in studying cell-penetrating peptides, hemolytic peptides, anticancer peptides, hormone proteins, autophagy proteins and Anti-CRISPR proteins, as well as other proteins, due to its high capability in identifying proteins. Protein identification.

With further improvement of protein sequence coding technologies and machine learning algorithms, sequence-based protein identification has emerged as a powerful tool for rapid protein identification. It is not only a tool to classify and identify certain proteins as predictive analyses of cell-penetrating peptides, hemolytic peptides, anti-cancer peptides, hormone proteins, autophagy proteins, and Anti-CRISPR proteins used the protein sequence encoding method combined with machine learning algorithms because of its precise of in protein identification studies. The importance is also on ascription and description of proteins by protein sequence coding and machine learning algorithms, which are very important in the strong analysis of cell-penetrating peptides [5], hemolytic peptides [6], anticancer peptides [7], hormone proteins [8], autophagy proteins [9] and Anti-Cancer Proteins [10] etc for trustworthy identification results. While influeca virus HA is central to its infection, advanced machine learning tools have lately focused on grouping viruses, detecting the host they attack, understanding mutations and changes to HA, knowing HA's structure and function and assessing the virus for whether it is highly infectious and transmissible used AI and machine learning with internet optimization [11-16]. There has been little research on applying machine learning techniques to HA detection through analyzing sequences.

Studies that use machine learning for HA have mainly focused on determining the subtype of the influenza virus, predicting the host, mutations and evolution, studying the function and architecture of HA and predicting pathogenicity and prevalence. However, HA is very important in the process of the influenza virus infecting us. No methods for recognizing HA using COVID-19 sequence data and ML algorithms are available today. We introduced a machine learning model in this work to successfully identify HA. Initially, we put together a collection of protein datasets called a benchmark dataset. Afterwards, we used feature extraction algorithms to encode the protein sequences. Following that, we collected all the features and applied a method called ANOVA with IFS to keep just the most significant features. Finally, this group of useful features formed the basis for the HA prediction model. This is how the process unfolds and you can see that in Figure 1.

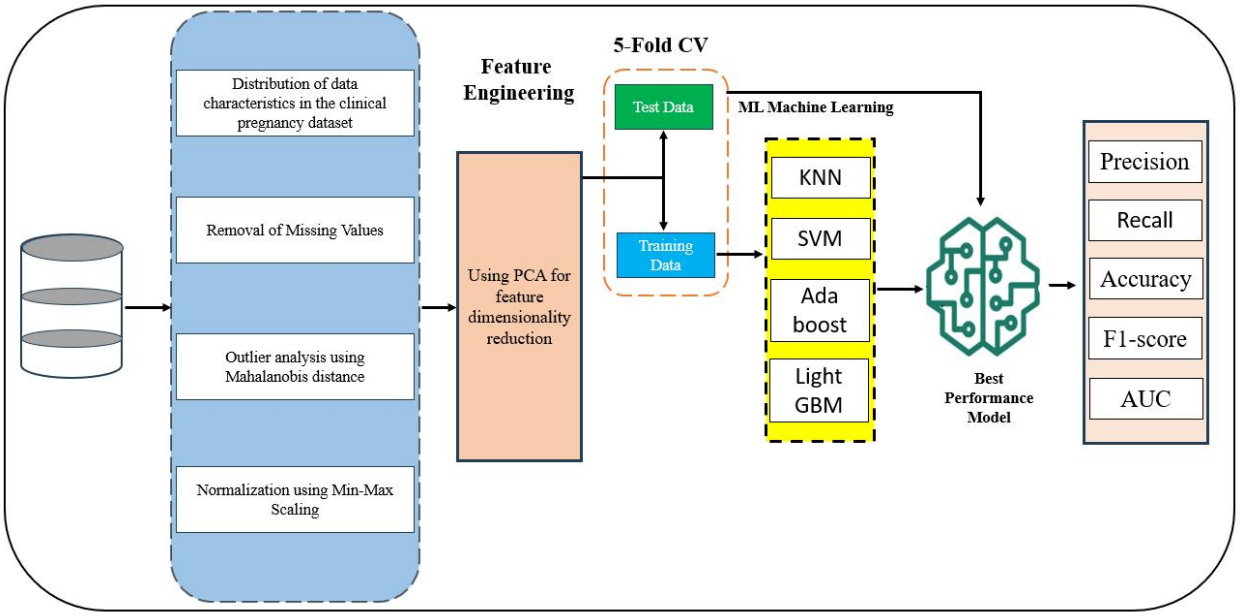

Figure 1: Proposed framework model of LightGBM hemagglutinin Proteins

2. Method and materials

2.1. Benchmark dataset

A benchmark set of data is indispensable for bioinformatics analysis [16- 17]. The information in this research was obtained from the Universal Protein Resource, referred to as UniProt [18]. Various processing steps were carried out to ensure the data was high quality. The program CD-HIT was used to filter out similar sequences among the retrieved ones[19]. A cut-off value of 80% similarity was set, and all sequences which showed >80% similarity were deleted. The non-HA dataset was subsampled to balance the positive and negative samples. The benchmark data contained 212 protein sequences, containing 106 HA sequences and 106 non-HA sequences. Data was distributed so that four parts went into the training set and one part went into the test set. You can access the model training and test set data at the link provided: https://github.com/Raho001. For testing, an extra file called 'test_data.txt' is available too. The different datasets have been thoroughly explained in Table 1. The Hemagglutinin protein was present in one sample, absent in another, but no reliable link was found. Positive and negative samples were obtained from test data and brought to the same condition.

Table 1: Accuracy (Acc) refers to the fraction of correctly categorized instances in the dataset.

	Total	Non-Redundancy	Training Data	Test Data	Training and testing
Hemagglutinin Protein	99	70%	85	20	106
Non-Hemagglutinin Protein	111	70%	84	22	106

2.2. Feature extraction

Protein identification and prediction depend on the process of feature extraction [20–21]. However, machine learning programs cannot by themselves analyze data from protein sequences for use in their models. Hence, it is important to transform proteins' sequences into numbers that machine learning programs can process. Such data can be changed into a numerical vector using different methods for feature extraction. A better choice of features will make classification more effective. EAAC [22], DDE [23] and PSSM are used to identify features from protein sequences. You can learn about these methods in the following sections.

A protein sequence P of length L may be denoted as:

$$P = R_1R_2R_3R_4R_5R_6 \cdots R_L \tag{1}$$

R1 represents the initial amino acid of the sequence, R2 signifies the subsequent amino acid, and so forth.

2.2.1. Dipeptide deviation from expected mean (DDE)

The DDE function vector comprises three principal parameters: theoretical mean value (**Tm**), theoretical difference (**Tv**), and dipeptide composition (Cc). **DC** (i), denotes the Cc of dipeptide i within peptide p.

Tm, Tv and Cc are the main parameters in the three-part DDE function vector. DC refers to Cc in dipeptide i in peptide p.

$$D_{c(i)} = \frac{n_i}{N} \tag{2}$$

Some studies have focused on extracting information (length samples) from 400 dipeptide features (which equals 20×20) to analyze specific amino acid combinations. The lengthis an accentuating property and describes how the samples are organized, but ineffective examples were thrown away during the process. Even so, I believe that both dipeptide 1 and N are L-1. So, we cannot consider it as another L-1 which is the expected amount in P. The letter TM (i) stands for the theoretical mean.

$$T_{M(i)} = \frac{C_{i1}}{C_N} \times \frac{C_{i2}}{C_N} \tag{3}$$

Every Ci1 and Ci2 indicates the number of codons and the amino acid connected to it. If the three termination codons are disregarded, the number of codons is still unchanged. Because TM (i) is distinct from peptide P, we were able to extract and forecast the properties of 400 dipeptides. In theory, TM (v) measures the variation seen in dipeptide I.

$$T_{v(i)} = \frac{T_{M(i)}\left(1 - T_{M(i)}\right)}{N} \tag{4}$$

The theoretical average value I, as per equation (2), is TM(i). Dipeptides are renumbered in peptide P, with N designated as L-1. Ultimately, DDE(i) is defined as

$$DDE_{(t)} = \frac{D_{c(i)} - T_{m(i)}}{\sqrt{T_{V(i)}}} \tag{5}$$

Finally, for every 400 features in length, the DDE vector score matrix is computed, utilizing feature vectors of 400 dimensions as follows:

$$DDE_p = \left\{DDE_{(i)}, \ldots, \ldots DDE_{(n)}\right\}, where, i = 1, 2, \ldots, 400 \tag{6}$$

2.2.2. Enhanced amino acid composition (EAAC)

This method is introduced by Chen et al.33. This algorithm extracts sequential protein information and calculates amino acid frequency as follows:

$$g(m,n) = \frac{H(m,n)}{H(n)}, m \in \{A, C, D, \cdots, Y\}, n \in \{W1, W2, \cdots WL\} \tag{7}$$

where H (m, n) is the quantity of amino acid type m, and H(n) represents the length of the n-th window.

2.2.3. Machine Learning Classifier

The researchers measure the accuracy of KNN, SVM, AdaBoost and **LightGBM** in handling different characteristics [24–25]. Once the parameters are tuned in the models, they undergo training with the same set of traits. By evaluating four algorithms using the Scikit-learn library, we found that the LightGBM classifier showed the best result in solving this particular problem [26].

2.2.4. Performance Evaluation

It is essential to decide on proper assessment criteria for a model's development. True positives, false negatives, true negatives and false positives are the major metrics used to evaluate models. True positives and true negatives are counted as TP and TN, respectively, while the model counts FN and FP for users who it mistook as positives or negatives. Using these four criteria, we decided on these six regularly used metrics to review how the model functions in this paper.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{8}$$

$$Sn = \frac{TP}{TP + FP} \tag{9}$$

$$Specificity = \frac{TN}{TN + FP} \tag{10}$$

$$MCC = \frac{TP \times TN - FP \times FN}{\sqrt{(TP + FN)(TP + FP)(TN + FP)(TN + FN)}} \tag{11}$$

Where TP, TN, FP, and FN are the true positives, true negatives, false positives and false negatives, respectively. We also applied receiver operating characteristic (ROC) curves to assess performance of the models. A higher AUC value indicates that the model is closer to 1 in a good model because it is designed based on the theory. You can compare the outcomes of the analysis and decide whether the methods you have used are appropriate and reliable with results in previous research. When an experiment is winding up, it's about checking, adding up, and comparing the data. A model developed for prediction utilising the DDE extraction method is presented.

3. Results

You can compare the findings of your research with those from previous research to decide if the methods you used are suitable and reliable. Completing an experiment mainly involves checking, calculating and comparing the data. We have created a model that uses the DDE extraction method for making predictions.

3.1 SHAP Feature Importance

Shapely Additive Explanations (SHAP) [27, 28] is a known method for studying the features of different protein samples. To figure out the importance of every feature, we take the mean of the absolute Shapley values for every feature by applying this mathematical expression to the data. Those traits with higher Shapley values are the most significant.

$$I_j = \frac{1}{n} \sum_{i=1}^{n} \left| \phi_j^{(j)} \right| \tag{12}$$

After being sorted by their importance, the features are then outlined with corresponding graphics. This figure illustrates the role of the SHAP feature in LightGBM-ACM for improving predictions about Hemagglutinin Protein involving phosphorylation sites. You can also use SHAP feature importance instead of permutation feature importance. But it is not possible to compare these value metrics as the method to assess permutation features relies on the model's performance drop. How large the feature is in the determination of SHAP is very important. All the attributes used to gather data were extracted from raw protein sequences. Next, we applied LightGBM method to identify the best set of feature scores. To determine how accurate the model is, we performed a 5-fold cross-validation as well as started testing it without prior exposure to the training data. At the end, we applied the LightGBM approach along with SHAP values to examine which features

are important and to interpret the model's results.

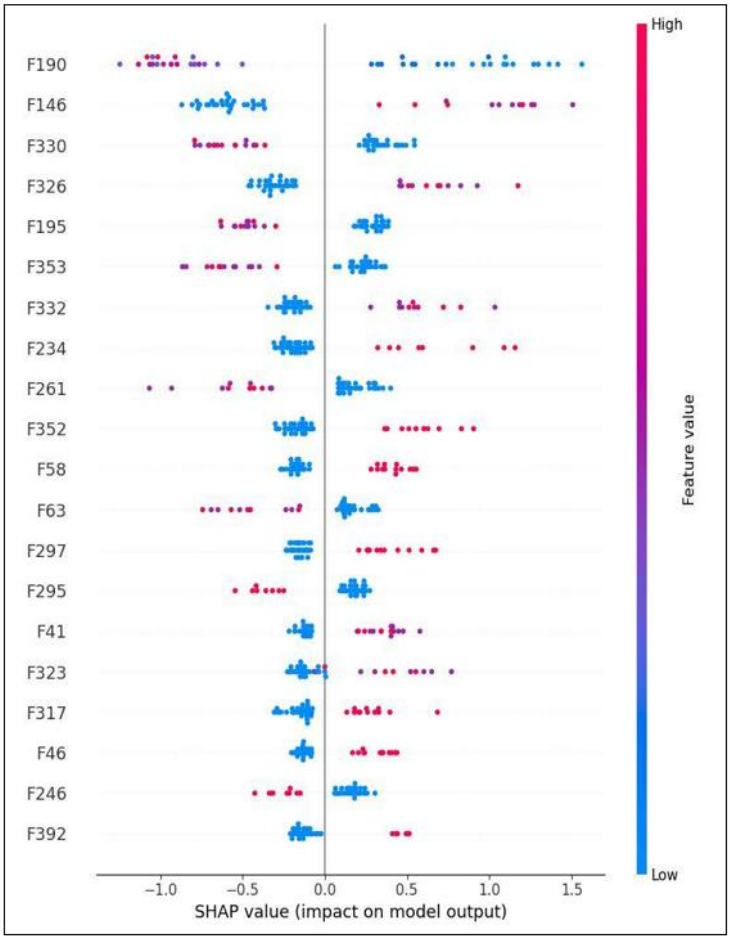

Figure 2: Evaluation of ranked attributes with the SHAP algorithm for the DDE model

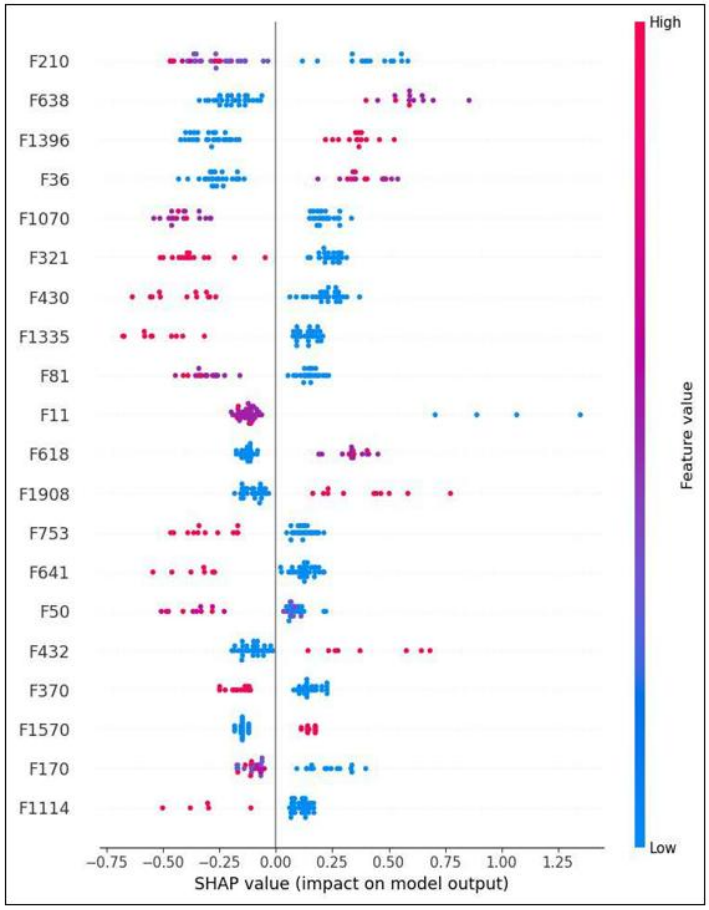

Figure 3: Assessment of ranked qualities with the SHAP method for the EAAC model.

The SHAP algorithm was applied to look at how significant each feature was in the DDE and EAAC models. Features F190 and F146 are highlighted in the DDE model for their strong and favorable influence on the predictions which shows up in the high SHAP value group concentrated in the positive area of the axis (from -1.0 to 1.5). Meanwhile, features F58 and F63 appeared to play different roles depending on the context, as their SHAP values ranged from positive to negative.

The EAAC model stood out because F210 and F638 were the most significant characteristics. However, the results from the SHAP analysis suggest that overall, the features in this model had little impact, shown by the small range of values (from -0.75 to 1.25). F36 and F1070 had somewhat different impacts, much like several parameters in the DDE model, but mostly mild.

3.2 Comparisons of metric performance

The suggested methods were evaluated by comparing the retrieved features to algorithms that help detect Hemagglutinin Protein. To compare the methods, I trained the models using GRU and applied two encoding techniques, DDE and EAAC. Using DDE-PSSM, the LightGBM algorithm was able to achieve an accuracy of 97.13%, precision of 100.0%, sensitivity of 94.29%, specificity of 100.0%, MCC of 94.55% and an AUC of 98.30%. The accuracy, precision, sensitivity, specificity, MCC and AUC values for using anti-hypertensives were 97.80%, 100.0%,

94.10%, 100.0%, 94.51% and 99.30%, respectively. The results displayed that LightGBM performed very well, mainly when it was combined with DEE-PSSM feature extraction as shown table 2. It was found that LightGBM performed better than K-NN, SVM and Adaboost which are considered traditional classifiers. To illustrate, lightGBM using DDE-PSSM had an accuracy of 98.30%, while RF and K-NN obtained accuracies of 84.0% and 81.96% each. Additionally, an analysis of ROC curves proved that LightGBM-based models perform better than others at every threshold. The results point out that by using a combination of DDE-PSSM, EAAC-PSSM and LightGBM, higher sensitivity and accuracy can be achieved when diagnosing the Hemagglutinin Protein.

Table 2. Varied metric performance contingent upon the test dataset for Hemagglutinin Protein

Cross-validation						
	Acc	Prec	Sens	Spec	Mcc	Auc
DDE-LightGBM	97.13%	100.00%	94.29%	100.00%	94.55%	98.30%
EAAC-LightGBM	97.08%	100.00%	94.10%	100.00%	94.51	99.30%

3.3 Comparative Examination of Diverse Classification Methods

The accuracy of the LightGBM-based DDE-PSSM profiles reached 98.30%, while that of the LightGBM-based profiles and the RF-based profiles was 77.06% and 92.26% respectively. Because of this, LightGBM performs more accurately than either Adaboost or KNN. The results for Hemagglutinin Protein, using the best set of features picked by the PSSM-EAAC strategy, are located in Figures 4 and 5. To achieve this score of 99.30%,ROC(AUC) with LightGBM and DDE or EAAC values were analyzed through comparisons.

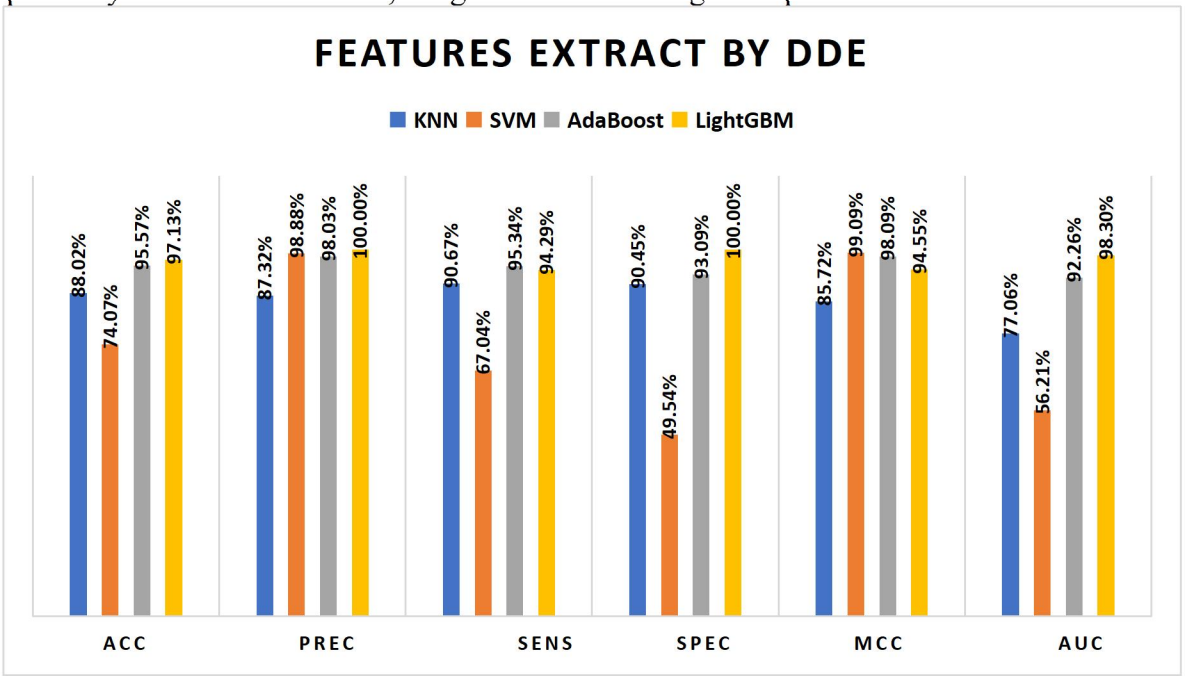

Figure 4: Compares the average classification accuracy (ACC) of different classifiers when using DDE-PSSM and PSSM-EAAC feature extraction methods. The results highlight performance variations across models based on the two representation strategies.

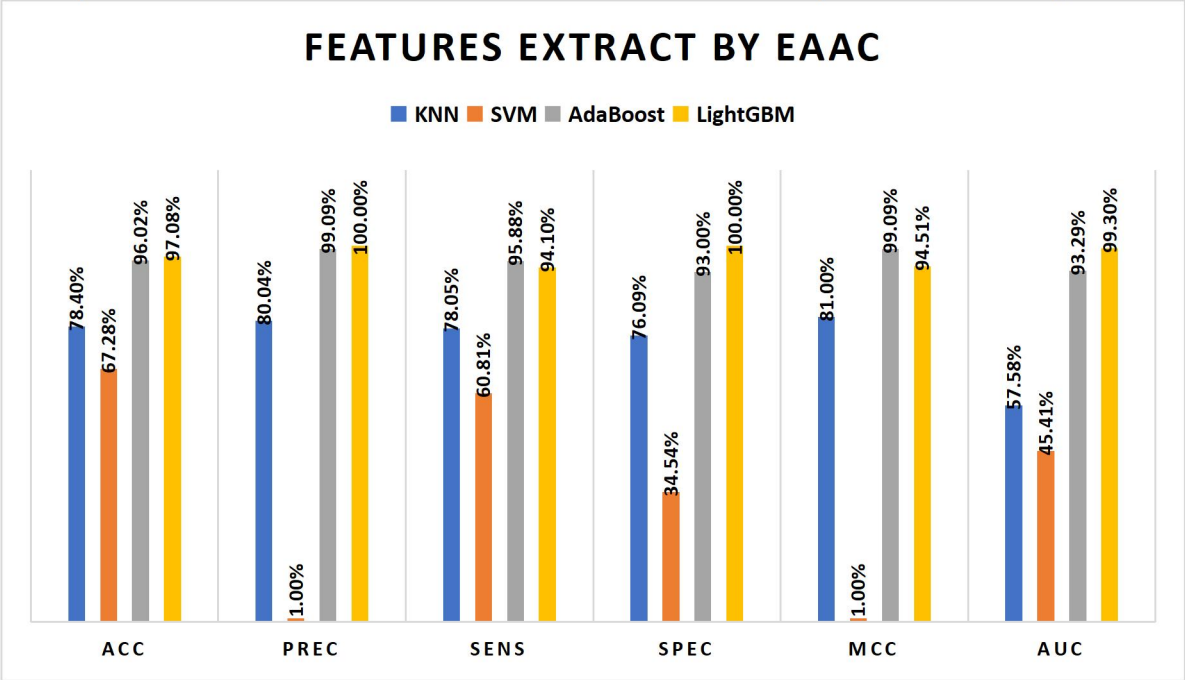

Figure 5: Presents the attributes obtained through the EAAC (Enhanced Amino Acid Composition) feature extraction approach.

Table 3. Performance metrics of PSSM-based models using the Hemagglutinin protein testing dataset.

Method	Classifiers	Acc	Prec	Sens	Spec	Mcc	Auc
DDE-PSSM	KNN	88.02%	87.32%	90.67%	90.45%	85.72%	77.06%
	SVM	74.07%	98.88%	67.04%	49.54%	99.09%	56.21%
	AdaBoost	95.57%	98.02%	95.34%	93.09%	98.09%	92.26%
	LightGBM	**97.13%**	**100.00%**	**94.29%**	**100.00%**	**94.55%**	**98.30%**
EAAC-PSSM	KNN	78.40%	80.04%	78.05%	76.09%	81.00%	57.58%
	SVM	67.28%	01.00%	60.81%	34.54%	01.00%	45.41%
	AdaBoost	96.02%	99.09%	95.88%	93.00%	99.09%	93.29%
	LightGBM	**97.08%**	**100.00%**	**94.10%**	**100.00%**	**94.51**	**99.30%**

The table 3. shows the comparison between different machine learning classifiers that were trained using the Hemagglutinin protein dataset and the two feature extraction techniques (DDE-PSSM and EAAC-PSSM). Among performance measures, we find Accuracy (Acc), Precision (Prec), Sensitivity (Sens), Specificity (Spec), Matthews Correlation Coefficient (Mcc) and Area Under the Curve (Auc). Regularly, LightGBM performs better than others, achieving perfect precision and specificity with both approaches to extracting features, together with a high accuracy of 97.08% and a high AUC of 99.30%. AdaBoost shows remarkable accuracy (95-13%) and precision (98-02%) in

processing protein data, suggesting it is robust. More often, DDE-PSSM results in higher accuracies compared to EAAC-PSSM, according to the performance of KNN and SVM. While the SVM performs well in terms of accuracy, it does not detect as many positive cases as negative cases. KNN does not perform as well as the other algorithms, but with DDE-PSSM features, its results are much better (88.02%), compared to EAAC-PSSM (78.40%).

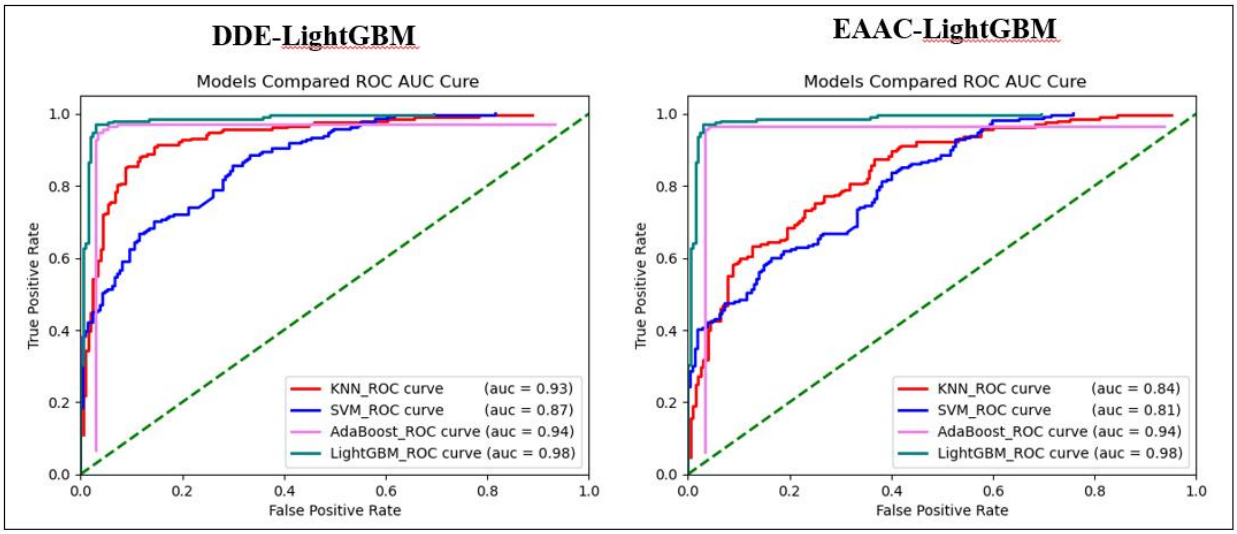

Figure 6: presents a comparative analysis of Receiver Operating Characteristic (ROC) curves evaluating the detection performance of Hemagglutinin (HA) proteins using the LightGBM-PSSM method.

3.4 Comparative analysis of Receiver Operating Characteristic (ROC) curves score

Four machine learning models, KNN, SVM, AdaBoost and LightGBM, were tested on two feature extraction approaches: DDE and EAAC, using this ROC curve research. Both the ROC curves and the Area Under the Curve (AUC) are included for each model to demonstrate how effective they are in telling apart positive cases from negative ones. All techniques use LightGBM as their classifier which outperforms the other two with nearly perfect AUC scores of 0.98. Comparing the algorithms demonstrates that DDE-based features have an edge over EAAC and usually perform better. To sum up, the comparison of classifiers is represented in Figure 6. The findings provide the results for various cutoffs and an initial evaluation of the ROC (AUC). In most cases, our LightGBM showed better results than other technologies.

Conclusion

Because hemagglutinin (HA) helps influence a vaccine, accurate identification of the protein is important. The paper describes a predictive model that works with HA protein sequence properties and was built using the Stacking method with the best set of features and basic classifiers. The findings suggest that our model is highly accurate and can apply its learning to new cases. The system is likely to provide dependable results in identifying and predicting HA events. In the future, more work will focus on finding additional ways to represent features and adjust the models to achieve even better results. Overall, our results highlight a good way to detect HA and support the creation of effective HA vaccines. As a result of these

contributions, influenza studies improve and future investigations in this domain are guided.

Reference:

1. Krammer F, Smith GJD, Fouchier RAM, Peiris M, Kedzierska K, Doherty PC, et al. Influenza. Nat Rev Dis Primers. (2018) 4:21. doi: 10.1038/s41572-018-0002-y
2. Uyeki TM, Hui DS, Zambon M, Wentworth DE, Monto AS. Influenza. Lancet. (2022) 400:693–706. doi: 10.1016/S0140-6736(22)00982-5
3. Byrd-Leotis, L., Cummings, R. D., & Steinhauer, D. A. (2017). The interplay between the host receptor and influenza virus hemagglutinin and neuraminidase. International journal of molecular sciences, 18(7), 1541.
4. Nuwarda RF, Alharbi AA, Kayser V. An overview of influenza viruses and vaccines. Vaccine. (2021) 9:27. doi: 10.3390/vaccines9091032
5. Win, T. S., Malik, A. A., Prachayasittikul, V., S Wikberg, J. E., Nantasenamat, C., & Shoombuatong, W. (2017). HemoPred: a web server for predicting the hemolytic activity of peptides. Future medicinal chemistry, 9(3), 275-291.
6. Schaduangrat, N., Nantasenamat, C., Prachayasittikul, V., & Shoombuatong, W. (2019). ACPred: a computational tool for the prediction and analysis of anticancer peptides. Molecules, 24(10), 1973.
7. Tang, H., Zhao, Y. W., Zou, P., Zhang, C. M., Chen, R., Huang, P., & Lin, H. (2018). HBPred: a tool to identify growth hormone-binding proteins. International Journal of Biological Sciences, 14(8), 957.
8. Yang, Y., Pan, Z., Sun, J., Welch, J., & Klionsky, D. J. (2024). Autophagy and machine learning: Unanswered questions. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease, 167263.
9. Zhu, L., Wang, X., Li, F., & Song, J. (2022). PreAcrs: a machine learning framework for identifying anti-CRISPR proteins. BMC bioinformatics, 23(1), 444.
0. Ghulam, A., Ali, F., Sikander, R., Ahmad, A., Ahmed, A., & Patil, S. (2022). ACP-2DCNN: Deep learning-based model for improving prediction of anticancer peptides using two-dimensional convolutional neural network. *Chemometrics and Intelligent Laboratory Systems*, *226*, 104589.
1. Ghulam, A., Sikander, R., & Ali, F. (2022). AI and Machine Learning-based practices in various domains: A Survey. *VAWKUM Transactions on Computer Sciences*, *10*(1), 21-41.
2. Ao C, Jiao S, Wang Y, Yu L, Zou Q. Biological sequence classification: A review on data and general methods. Research. (2022) 2022:0011. doi: 10.34133/research.0011
3. Xu Y, Wojtczak D. Dive into machine learning algorithms for influenza virus host prediction with hemagglutinin sequences. Biosystems. (2022) 220:104740. doi: 10.1016/j. biosystems.2022.104740
4. Jones, S., Nelson-Sathi, S., Wang, Y., Prasad, R., Rayen, S., Nandel, V., ... & Pillai, R. M. (2019). Evolutionary, genetic, structural characterization and its functional implications for the influenza A (H1N1) infection outbreak in India from 2009 to 2017. Scientific reports, 9(1), 14690.
5. Rodríguez, J. M., Timm, D. E., Titus, G. P., Beltrán-Valero de Bernabé, D., Criado, O., Mueller, H. A., ... & Penalva, M. A. (2000). Structural and functional analysis of mutations in alkaptonuria. Human molecular genetics, 9(15), 2341-2350.
6. Rahu, Ghulam, A., Mansoor Hyder Depar, Sher Muhammad Daudpoto, Mir Sajjad Hussain Talpur, Muhammad Malook Rind, and Gordhan Das. "EXTENDING OFF-PAGE SEARCH ENGINE OPTIMIZATION (SEO) TECHNIQUES BASED ON GOOGLE SEO TECHNIQUES MODEL." *Science International* 28, no. 5 (2016).

17. Wei Y, Zou Q, Tang F, Yu L. WMSA: a novel method for multiple sequence alignment of DNA sequences. Bioinformatics. (2022) 38:5019–25. doi: 10.1093/bioinformatics/btac658
18. Sureyya Rifaioglu, A., Doğan, T., Jesus Martin, M., Cetin-Atalay, R., & Atalay, V. (2019). DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks. Scientific reports, 9(1), 7344.
19. Fu LM, Niu BF, Zhu ZW, Wu ST, Li WZ. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. (2012) 28:3150–2. doi: 10.1093/bioinformatics/bts565
20. Ghulam, Ali, Tarique Ali, Taha Hussain, Nida Jabeen, Taiyaba Qureshi, Mujeeb ur Rehman, Rahu Sikander, and Sultan Ahmed. "SOCIAL SCIENCE REVIEW ARCHIVES ISSN Print: 3006-4694."
21. Saravanan V, Gautham N. Harnessing Computational Biology for Exact Linear B-CellEpitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. OMICS.2015; 19(10):648-658. doi:10.1089/omi.2015.0095
22. Sorkhi, A. G., Pirgazi, J., & Ghasemi, V. (2022). A hybrid feature extraction scheme for efficient malonylation site prediction. Scientific Reports, 12(1), 5756.
23. Sikander, R., Wang, Y., Ghulam, A., & Wu, X. (2021). Identification of enzymes-specific protein domain based on DDE, and convolutional neural network. Frontiers in Genetics, 12, 759384.
24. Sikander, R., Rehman, M., Brohi, T. A., Ahmed, A., Ghulam, A., & Ahmed, S. (2024). An Explainable Identifier of iGHBPs Peptides Based on Deep PSSM Features and Learning Approaches. Insights-Journal of Health and Rehabilitation, 2(2 (Health & Allied)), 571-579.
25. Khorshid, M., Abou-El-Enien, T. H., & Soliman, G. (2015). A comparison among support vector machine and other machine learning classification algorithms. IPASJ International Journal of Computer Science (IIJCS), 3(5).
6. Yang, H., Chen, Z., Yang, H., & Tian, M. (2023). Predicting coronary heart disease using an improved LightGBM model: Performance analysis and comparison. IEEE Access, 11, 23366-23380.
7. Saravanan V, Gautham N. Harnessing Computational Biology for Exact Linear B-CellEpitope Prediction: A Novel Amino Acid Composition-Based Feature Descriptor. OMICS.2015; 19(10):648-658. doi:10.1089/omi.2015.0095
8. Saravanan, V., Gautham, N. BCIgEPRED—a Dual-Layer Approach for Predicting LinearIgE Epitopes. Mol Biol 52, 285–293 (2018).