

Prediction of Human Papillomavirus-Host Oncoprotein Interactions Using Deep Learning

Sheila Santa^{1,2}, Samuel Kojo Kwofie³, Kwasi Agyenkwa-Mawuli⁴, Osbourne Quayle¹, Charles A Brown² and Emmanuel A Tagoe²

¹Department of Biochemistry, Cell & Molecular Biology/West African Centre for Cell Biology of Infectious Pathogens (WACCBIP), College of Basic and Applied Sciences, University of Ghana, Accra, Ghana. ²Department of Medical Laboratory Sciences, School of Biomedical and Allied Health Sciences, College of Health Sciences, University of Ghana, Accra, Ghana. ³Department of Biomedical Engineering, School of Engineering Sciences, College of Basic and Applied Sciences, University of Ghana, Accra, Ghana. ⁴Noguchi Memorial Institute for Medical Research, College of Health Sciences, University of Ghana, Accra, Ghana.

Bioinformatics and Biology Insights
Volume 18: 1–9
© The Author(s) 2024
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/11779322241304666



ABSTRACT

BACKGROUND: Human papillomavirus (HPV) causes disease through complex interactions between viral and host proteins, with the PI3K signaling pathway playing a key role. Proteins like AKT, IQGAP1, and MMP16 are involved in HPV-related cancer development. Traditional methods for studying protein-protein interactions (PPIs) are labor-intensive and time-consuming. Computational models are becoming more popular as they are less labor-intensive and often more efficient. This study aimed to develop a deep learning model to predict interactions between HPV and host proteins.

METHOD: To achieve this, available HPV and host protein interaction data was retrieved from the protocol of Eckhardt et al and used to train a Recurrent Neural Network algorithm. Training of the model was performed on the SPYDER (scientific python development environment) platform using python libraries; Scikit-learn, Pandas, NumPy, and TensorFlow. The data was split into training, validation, and testing sets in the ratio 7:1:2, respectively. After the training and validation, the model was then used to predict the possible interactions between HPV 31 and 18 E6 and E7, and host oncoproteins AKT, IQGAP1 and MMP16.

RESULTS: The model showed good performance, with an MCC score of 0.7937 and all other metrics above 88%. The model predicted an interaction between E6 and E7 of both HPV types with AKT, while only HPV31 E7 was shown to interact with IQGAP1 and MMP16 with confidence scores of 0.9638 and 0.5793, respectively.

CONCLUSION: The current model strongly predicted HPVs E6 and E7 interactions with PI3K pathway, and the viral proteins may be involved in AKT activation, driving HPV-associated cancers. This model supports the robust prediction of interactomes for experimental validation.

KEYWORDS: Human papillomavirus, protein-protein interactions, recursive neural network, deep learning, machine learning, PI3K pathway and oncoproteins

RECEIVED: January 31, 2024. **ACCEPTED:** November 16, 2024.

TYPE: Method and Protocol

FUNDING: The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: Sheila Santa was supported by a WACCBIP-World Bank ACE PhD fellowship (WACCBIP + NCDs: Awandare)* and a DELTAS Africa grant (DEL-15-007: Awandare). The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS)'s Alliance for Accelerating Excellence in Science in Africa and supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust (107755/Z/15/Z: Awandare) and the UK government. The views expressed in this publication are those of the author(s) and not necessarily those of AAS, NEPAD Agency, Wellcome Trust, or the UK government.

DECLARATION OF CONFLICTING INTERESTS: The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

CORRESPONDING AUTHOR: Samuel Kojo Kwofie, Department of Biomedical Engineering, School of Engineering Sciences, College of Basic and Applied Sciences, University of Ghana, Accra, Ghana. Email: skkwofie@ug.edu.gh

Emmanuel a Tagoe, Department of Medical Laboratory Sciences, School of Biomedical And Allied Health Sciences, College of Health Sciences, University of Ghana, Accra, Ghana. Email: eatagoe@ug.edu.gh

Introduction

Persistent infection with high-risk human papillomavirus (HPV) has been implicated in approximately 600 000 cases of cancers of the cervix, oropharynx, anus, vulvovaginal, and penis.^{1–3} The genome of HPV is divided into 3 parts. The first part is the early region containing the early genes. The early genes E1 and E2 code for the major replication proteins E1 and E2 while E4 and E5 proteins aid in genome amplification. The E6 and E7 proteins are the main oncoproteins that drive carcinogenesis. The second is the late region, containing the late genes L1 and L2, which encode the L1 major and L2 minor capsid proteins. The third part is the long control region

(LCR) which is the only noncoding region of the genome, and contains the early viral promoters, enhancers, and the origin of viral DNA replication.^{4,5}

The ability of HPV to cause disease relies on a complex interaction between early viral proteins and host proteins. The carcinogenic transformation of HPV-associated lesions is as a result of deregulation of virus-host cross-talk, leading to over-expression of E6 and E7 viral oncogenes and consequently the accumulation of cellular genetic mutations.⁶ The PI3K/Akt signaling pathway has been reported to play a central role in the virus/host cell cross-talk of HPV-positive cancer cells.⁷ The AKT, IQGAP1 and MMP16 have all been implicated in



HPV-associated carcinogenesis through the PI3K pathway.⁸⁻¹⁰ Therefore, understanding how HPV proteins and host proteins interact in the disease state will be very useful in elucidating infection mechanisms and unraveling more efficient approaches to managing the disease. This can be done through prevention, risk identification based on genetic profiles, and better drug design. Currently, available experimental approaches for studying protein-protein interactions (PPI) include pull down, 2-hybrid, gel filtration chromatography, isothermal titration calorimetry, Förster resonance energy transfer, luminescent oxygen channeling, reflectometric interference spectroscopy, and other high-throughput biological techniques.¹¹ Even though these experimental methods have been used to identify PPIs in several model organisms, they are time-consuming and labor-intensive. In addition, their applicability depends on the effectiveness of assay protocols as some may not work on certain classes of proteins. Similarly, experimental methods may miss weak interactions and leave out many transient interactions.¹²

Therefore, the use of computational methods to complement experimental methods in predicting PPIs is gaining widespread acceptance as they are less labor-intensive, fast and more efficient. Machine learning (ML), a subfield of artificial intelligence that focuses on using data to learn associations and make predictions, is the most common computational method in use.¹³ Generally, ML algorithms have 4 categories: supervised learning, unsupervised learning, semi-supervised learning, and reinforcement learning.¹⁴ Among these methods, supervised learning, which is the focus of this study, is the most used approach. The supervised ML method relevant to PPI predictions is the binary classification approach. This involves the use of binary labeled training data (2 input proteins and their interactions), to construct a model that infers labels for test data (which usually originates from the same distribution as the training data, but without labels).¹⁵ The model then extracts the relevant features from each of the protein representations and uses them for PPI prediction. In supervised ML, training data is used to adjust model parameter settings that tend to predict known interactors accurately.¹⁵ Once trained, the model is expected to predict labels in the training data accurately. The test data, however, serves as unseen data and provides information on how well the model performs beyond the training data. This helps to test the model's ability to generalize on new data sets. The aim of this study was to adopt a suitable deep learning model to predict HPV and host PPIs.

Materials and Methods

The main data set used for this experiment was from the protocol of Eckhardt et al. They mapped the global network of HPV31 proteins and host PPIs by purifying the complete set of HPV31 proteins in multiple cell lines, followed by mass

spectrometry analysis.¹⁶ The data sets had 3 main subsets with interactions between HPV31 and host protein pairs: HEK293 PPIs (405 positives and 393 negatives), C33A PPIs (137 positives and 128 negatives), and Het-1A PPIs (84 positives). The positive interactions were defined in the experiments as those with MiST scores greater than 0.75 on a scale of 0 to 1. The negative interactions were sampled out of the rest as those with MiST scores less than 0.02. The total data set used for the model training, validation and testing was the combined data sets of HEK292 and C33A (1064 PPIs). Out of this, the training set had 744 PPIs, the validation set had 106 PPIs and the test set had 214 PPIs. The Het-1A data set was held as an external validation data set while additional data on HPV18 and host PPIs was obtained from the Viruses.STRING online database,¹⁷ to test the model's ability to predict other HPV types and host PPIs.

Methods

Obtaining data. The amino acid sequences of the proteins for each interaction in the data set were obtained from UniProt database as FASTA files¹⁸ using a custom python script and a spreadsheet containing their unique UniProt accession numbers. The files for each protein pair were then stored in a working directory.

Preprocessing data. The retrieved data was preprocessed by converting each protein sequence into a suitable data format for the prediction model, using a technique known as tokenization. This was based on the approach used in a similar study,¹⁹ where the protein sequences were broken into 3-gram non-overlapping tokens. The attributes used to describe each PPI data set consist of the labels, Protein_1_bait and Protein_2_pre, which were assigned to the HPV and human host protein pair, respectively, while 0 and 1 were assigned to the pair for interactivity (1 for interaction, 0 for non-interaction). The amino acid sequences were retrieved and modeled as a "3-gram" sequence, where three contiguous amino acids were joined to form a single unit or "word" as done in AI-based natural language processing (NLP). For example, an amino acid "X" preceded by amino acid "Y," and followed by amino acid "Z," together form a triplet, "YXZ." Every triplet in the sequence is then encoded as an integer (1,2,3, . . . [last token]) to produce a series of tokens. To obtain a fixed length of the tokenized protein sequences (1000 elements), the token "0" was reserved for padding the length of the numeric sequences to the left of the vector until the length of 1000 was obtained. If the vector sequence of the protein was longer than 1000, it was truncated, discarding the leftmost elements (the beginning of the amino acid chain) in the sequence (Figure 1). A 3-gram modeling approach performed well in similar instances whereby n-gram modeling in protein informatics was applied using NLP.^{19,20} Nevertheless, reported n values for n-gram modeling can range from n=2 up to n=6.^{21,22}

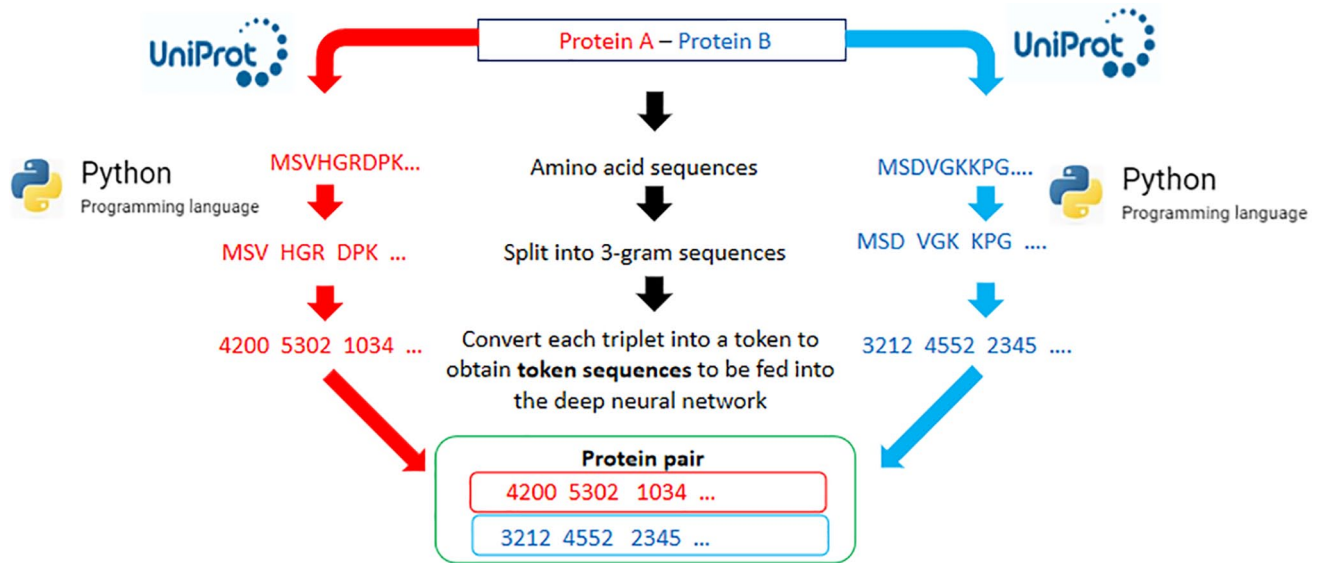


Figure 1. The data preprocessing workflow for training the deep learning model. The protein sequences are obtained from the UniProt online repository and then preprocessed using python programming to obtain tokenized sequences of numerical vectors.

Training. Model training was performed on the SPYDER platform using python libraries: Scikit-learn, Pandas, NumPy, and TensorFlow. The data was split into training, validation, and testing sets in the ratio 7:1:2, respectively. The deep learning algorithm chosen was the Recurrent Neural Network (RNN) because of its inherent feature extraction capabilities which is suitable for similar applications in previous ML PPI studies.^{19,23} The RNN comprises a deep learning structure that uses past information to improve the performance of the network on current and future inputs.²⁴ The training set of the data was used to train the network while the validation set was used to test the model's performance at the end of each epoch (training cycle). Hyper-parameters (data weights, number of dense layers, activation functions, and loss functions) were adjusted to improve this validation accuracy and reduce validation loss until they did not change significantly after a given number of epochs. The model was fed with the tokenized sequences of 2 proteins, Protein_1_bait and Protein_1_pre, representing viral and host proteins, respectively, with their label (whether the protein pair interacts or not), as its inputs. The 2 proteins were processed in 2 separate branches of the network, where the features that describe the proteins were learned. Each branch comprised an embedding layer, a recurrent layer (with GRU units) and a fully connected layer to rearrange and recombine the information extracted by the 2 previous layers. The features that each extractor branch had computed were then merged into a single vector by concatenation, and the fully connected layer was used to combine the features of both proteins. The output was then interpreted as the probability of a sample (a protein pair) belonging to one of 2 classes (interaction, $\geq 50\%$ or no interaction $< 50\%$). The

final label was then assigned based on the score criterion. The parameters of the network were optimized by comparing the predicted label to the true label to improve the performance (Figure 2).

Evaluation. After the training phase, the performance of the model was measured on the test data set using standard statistical metrics comprising balanced accuracy, precision, recall, specificity, Matthew's Correlation Coefficient (MCC), F-1 score, and Area Under the Receiver Operating Characteristic Curve (ROC-AUC).^{25,26} These metrics are defined by the equations below:

$$\text{Balanced accuracy} = \frac{\text{Sensitivity} + \text{Specificity}}{2}$$

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

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

$$\text{F1 Score} = \frac{TP}{TP + \frac{1}{2} (FP + FN)}$$

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

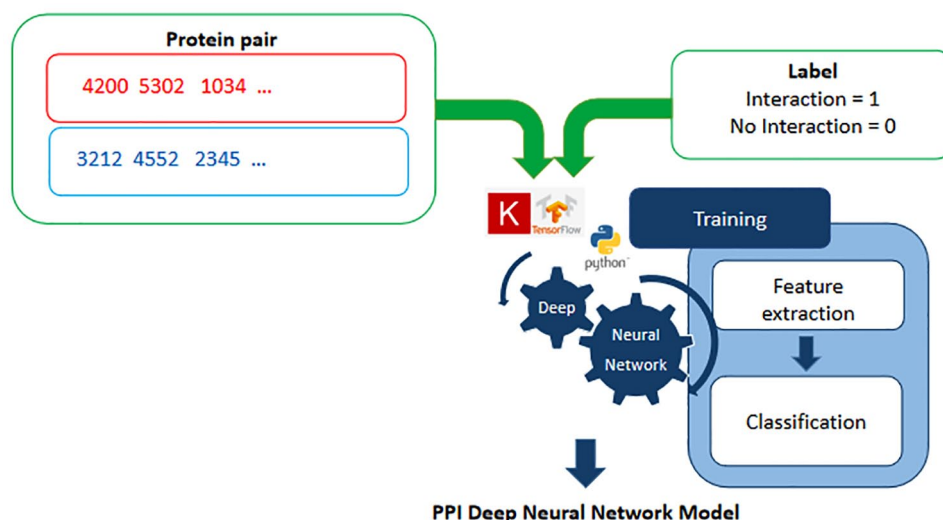


Figure 2. A schema summarizing the training of the neural network. The tokenized sequences in a given pair and their interaction label are fed into the neural network whose architecture is built to carry out feature extraction and then classification to obtain a model that can make predictions on new data.

where TP, TN, FP, and FN represent the numbers of true-positive, true-negative, false-positive, and false-negative samples, respectively, based on the confusion matrix for classifications.

The model was further validated with 2 data sets: the Het-1A data set on HPV31 interactions, and a data set on HPV18 and host PPIs from Viruses.STRING.¹⁷ The latter was used to investigate the predictive power of the trained model on HPV18 interactions despite having only trained on HPV31 interactions. The Het-1A data set of 84 positive interactions was pre-filtered against the combined HEK295 and C33A data set to eliminate common interactions and hence remove “familiar” contamination to it as external validation set. This dropped the positive interactions in Het-1A from 84 to 33. Tables 1 to 4 in the “Results” section summarize this.

Results

Evaluation of model performance

The values of the various metrics scored by the model on the test set (except for MCC) are provided (Figure 3). The MCC score was 0.7937, which indicated that the model was a good classifier.

Generally, the model showed good performance, with all parameters showing performance above 88% which comparatively is not far off from the performance of models trained elsewhere.^{19,26}

Validation of model

Out of a total of 33 HPV31 and host protein positive interactions from Het-1A used to validate the model, the model was

able to correctly predict 25 as positive interactions (true positives), while 8 were predicted as negative interactions (false negatives). This gave the model a 75.75% recall on the Het-1A data set (Table 1).

The model was further tested to determine its ability to predict interaction between host proteins and proteins of a different HPV type (HPV18). A data set of 10 positive HPV18 and host PPIs was retrieved from Viruses.STRING database¹⁷ and used to test the performance of the model. The model was able to correctly predict 7 out of the 10 positive interactions, giving it a recall of 70% on the HPV18 data set. The results for HPV18 and host protein interaction prediction by the model are shown in Table 2.

Model predictions

The model was then used to predict the possible interaction between HPV E6 and E7 and host oncoproteins AKT, IQGAP1, and MMP16. This was done for both HPV18 and HPV31, which gave quite similar results in predictions with only slight variations in probabilities. Table 3 presents the results for HPV31, while Table 4 shows those of HPV18.

Subsequently, HPV31 E6 and E7 were predicted to interact with AKT, while only HPV31 E7 was shown to interact with IQGAP1 and MMP16. Similarly, when the model was used to predict the possible interactions between HPV18 E6 and E7 and AKT, IQGAP1, and MMP16, the same predictions were made except with slight differences in confidence scores. This correlation points to shared similarities in the interactions HPV18 and HPV31 have with host oncoproteins AKT, IQGAP1, and MMP16. The resources including scripts used

Table 1. Validation of performance of model using external data set showing the MiST and PPI model scores as well as the assigned labels.

PROTEIN_1_BAIT	UNIPROT_ACCESSION_NUMBER_1	PROTEIN_2_PREY	UNIPROT_ACCESSION_NUMBER_2	MIST SCORE	ORIGINAL LABEL	PPI_MODEL SCORE	MODEL ASSIGNED LABEL
HPV31L2_S	P17389	SQRD_HUMAN	Q9Y6N5	0.9901	1	0.835511923	1
HPV31L2_S	P17389	SPTC1_HUMAN	O15269	0.9901	1	0.078733124	0
HPV31L1_S	P17388	HNRL1_HUMAN	Q9BUJ2	0.98776	1	0.34279865	0
HPV31L2_S	P17389	ERLN1_HUMAN	O75477	0.98774	1	0.94157517	1
HPV31L1_S	P17388	RFOX2_HUMAN	O43251	0.98651	1	0.412922651	0
HPV31L1_S	P17388	ZN326_HUMAN	Q5BKZ1	0.9802	1	0.903198719	1
HPV31L1_S	P17388	WIBG_HUMAN	Q9BRP8	0.96714	1	0.995505869	1
HPV31L1_S	P17388	NOP58_HUMAN	Q9Y2X3	0.96416	1	0.656340361	1
HPV31L2_S	P17389	CMC1_HUMAN	O75746	0.95817	1	0.934573591	1
HPV31L2_S	P17389	TBB8_HUMAN	Q3ZCM7	0.91231	1	0.22585085	0
HPV31L1_S	P17388	FBLN3_HUMAN	Q12805	0.89242	1	0.868786156	1
HPV31L1_S	P17388	RL10A_HUMAN	P62906	0.88017	1	0.894154668	1
HPV31L1_S	P17388	CHD4_HUMAN	Q14839	0.88014	1	0.966781557	1
HPV31E7_S	P17387	CCNA2_HUMAN	P20248	0.88006	1	0.986383915	1
HPV31L1_S	P17388	HNRH2_HUMAN	P55795	0.88006	1	0.197558418	0
HPV31L1_S	P17388	KAP0_HUMAN	P10644	0.88	1	0.566887438	1
HPV31L1_S	P17388	HNRPL_HUMAN	P14866	0.87989	1	0.71629715	1
HPV31L2_S	P17389	PLK1_HUMAN	P53350	0.87829	1	0.967017055	1
HPV31L1_S	P17388	IMA7_HUMAN	O60684	0.87689	1	0.529450834	1
HPV31L1_S	P17388	CKAP2_HUMAN	Q8WWK9	0.87689	1	0.271408409	0
HPV31L1_S	P17388	DDX27_HUMAN	Q96GQ7	0.87689	1	0.846727788	1
HPV31L1_S	P17388	DRG1_HUMAN	Q9Y295	0.87673	1	0.852020621	1
HPV31L1_S	P17388	TCOF_HUMAN	Q13428	0.87673	1	0.931821406	1
HPV31L1_S	P17388	H11_HUMAN	Q02539	0.87541	1	0.071080804	0
HPV31L1_S	P17388	NP1L4_HUMAN	Q99733	0.87525	1	0.881988466	1
HPV31L1_S	P17388	ABCF1_HUMAN	Q8NE71	0.87504	1	0.390177667	0
HPV31E2_S	P17383	SMC6_HUMAN	Q96SB8	0.86954	1	0.892946899	1
HPV31L1_S	P17388	RFC1_HUMAN	P35251	0.8672	1	0.95859158	1
HPV31L2_S	P17389	TIM50_HUMAN	Q3ZCQ8	0.85303	1	0.548475564	1
HPV31L1_S	P17388	DHX36_HUMAN	Q9H2U1	0.84899	1	0.607559264	1
HPV31L1_S	P17388	HNRL2_HUMAN	Q1KMD3	0.82506	1	0.626150787	1
HPV31L2_S	P17389	KANK2_HUMAN	Q63ZY3	0.81256	1	0.530457973	1
HPV31L2_S	P17389	C1QBP_HUMAN	Q07021	0.76217	1	0.756749749	1

Table 2. Results for HPV18 and host protein interaction prediction by the model comprising confidence and PPI model scores.

PROTEIN_1_BAIT	UNIPROT_ACCESSION_NUMBER_1	PROTEIN_2_PREY	UNIPROT_ACCESSION_NUMBER_2	CONFIDENCE SCORE	LABEL	PPI_MODEL SCORE	ASSIGNED LABEL
HPV-18_E6	P06463	TP53	Q12888	0.966	1	0.843127787	1
HPV-18_E7	P06788	CSNK2A1	P68400	0.882	1	0.514482141	1
HPV-18_E5	Q549H4	CTCF	P49711	0.8	1	0.254359514	0
HPV-18_E7	P06788	EEF1A1	P68104	0.8	1	0.899078369	1
HPV-18_E7	P06788	HDAC9	Q9UKV0	0.8	1	0.974592984	1
HPV-18_E6	P06463	DLG1	Q12959	0.699	1	0.515228689	1
HPV-18_E6	P06463	DLG4	P78352	0.695	1	0.417945951	0
HPV-18_E6	P06463	CASK	P07498	0.648	1	0.902936876	1
HPV-18_E6	P06463	TERT	O14746	0.647	1	0.403731585	0
HPV-18_E2	P06790	CDC25A	P30304	0.627	1	0.522622943	1

Table 3. HPV31 E6 and E7 interaction with AKT, IQGAP1, and MMP16 with PPI_model predictions and confidence scores.

PROTEIN_1_BAIT	UNIPROT_ACCESSION_NUMBER_1	PROTEIN_2_PREY	UNIPROT_ACCESSION_NUMBER_2	PPI_MODEL PREDICTION	CONFIDENCE SCORE
HPV-31_E6	P17386	AKT	Q12888	1	0.882792413
HPV-31_E6	P17386	IQGAP1	P68400	0	0.442386329
HPV-31_E6	P17386	MMP16	P49711	0	0.247584611
HPV-31_E7	P17387	AKT	P68104	1	0.860466361
HPV-31_E7	P17387	IQGAP1	Q9UKV0	1	0.963706136
HPV-31_E7	P17387	MMP16	Q12959	1	0.57931447

Table 4. HPV18 E6 and E7 interaction with AKT, IQGAP1, and MMP16 showing the confidence scores and PPI_model predictions.

PROTEIN_1_BAIT	UNIPROT_ACCESSION_NUMBER_1	PROTEIN_2_PREY	UNIPROT_ACCESSION_NUMBER_2	PPI_MODEL PREDICTION	CONFIDENCE SCORE
HPV-18_E6	P06463	AKT	Q12888	1	0.843128
HPV-18_E6	P06463	IQGAP1	P68400	0	0.361482
HPV-18_E6	P06463	MMP16	P49711	0	0.190157
HPV-18_E7	P06788	AKT	P68104	1	0.899078
HPV-18_E7	P06788	IQGAP1	Q9UKV0	1	0.974593
HPV-18_E7	P06788	MMP16	Q12959	1	0.665481

have been deposited in a GitHub repository (Access link: Supplementary).

Discussion

This study presents a model for predicting HPV and host PPIs using deep learning-based RNN algorithm. In this study, we set out to design a neural network to predict HPV and host

protein interactions using the primary amino acid sequences. To achieve this, we utilized already available data,¹⁶ thus circumventing the need for feature engineering associated with prediction problems, and then transformed the data to an RNN architecture selection problem. The resulting model is the first trained on this data set. The performance of our model was evaluated by criteria such as balanced accuracy, recall, precision,

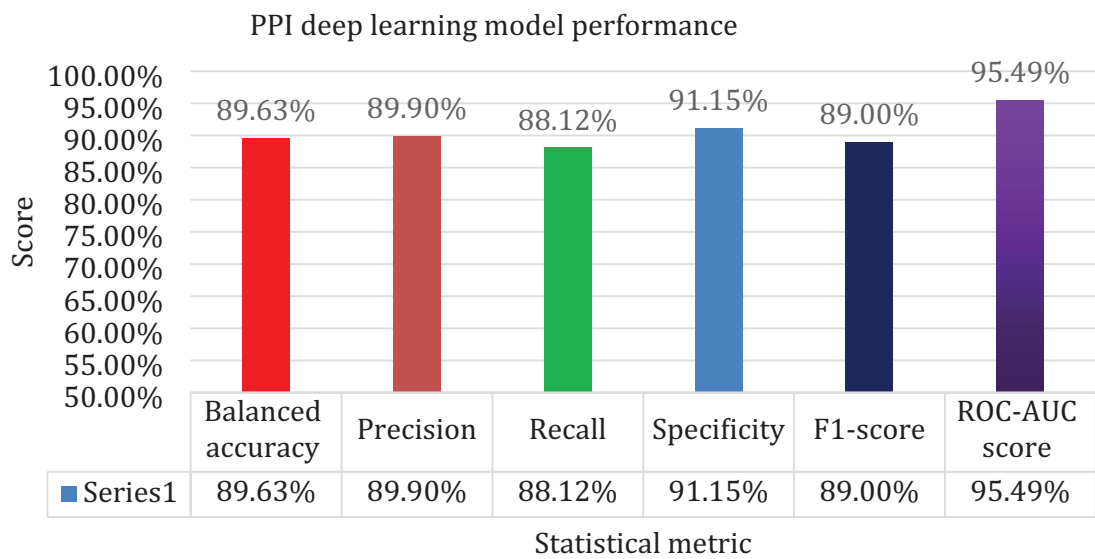


Figure 3. Model's performance on the evaluation metrics. All the metrics scored above 88% indicating the model's strong performance.

F-score, and MCC. Since PPI prediction is mostly aimed at correctly predicting the interacting protein pairs, the sensitivity and accuracy indicators were used to assess the model's ability to predict positive data. Similarly, the MCC was used to evaluate the reliability and stability of the model when dealing with unbalanced data, and the receiver operating characteristic (ROC) was used to appraise the performance of a set of classification results while the area under ROC (AUC) was computed as an important evaluation indicator.²⁷

Our model showed good performance, with all the evaluation matrices ranging from 88% to 95%. When the performance of our model was compared to the performances of similar deep learning models,^{26,27} we noticed that the evaluation matrices of most of the models were slightly above the performance of our model (above 90%). For example, the performance of the deep learning model developed previously²⁷ on human, *Helicobacter pylori*, and *Saccharomyces cerevisiae* data sets had accuracy around 98% and sensitivity, precision, and MCC around 98.47%, 98.67%, and 97.19%, respectively.²⁷ Similarly, the performance of a deep neural network²⁶ on a combined data set comprising human, *E. coli*, *Drosophila*, *C. elegans*, and *Mus musculus* data sets achieved accuracy of 0.9941, recall of 0.9963, precision of 0.9915, F-score of 0.9939, and MCC of 0.9883.²³ The difference in performance observed between our model and other similar deep learning models could be due to the size of the data set used in the training of the models. For example, the data size used for the design of our model was approximately 1000 protein pairs, mainly because of paucity of data on HPV and host PPIs; while the data size of Wang et al, and Li et al, was approximately 70 000 and 60 000 protein pairs, respectively. Hence our data space may not have covered enough samples with intrinsic features. For instance, in Table 1, row 2, the interaction between HPV31L2_Sand SPTC1_HUMAN was the highest with an

MiST score of 0.9901, and yet the model prediction scored 0.0787. This study used a 3-gram tokenization, which had been shown to produce robust outcomes. In future, once input data size increases considerably, the n-gram can be varied to evaluate its impact on the performance metrics. Similarly, the model can be optimized to learn features and increase its performance.

Other supervised ML models for PPI data sets have been trained to make predictions; however, they have been mainly based on the SVM algorithm. Five SVM models were trained on 16000 diverse PPI pairs²⁸ and the sequences were characterized by a conjoint triad descriptor and were used to train the models using a custom S-kernel function. They obtained accuracies > 82.75%, sensitivity values > 84.00% and precision scores > 82.75%. Similar models were trained,²⁹ however, the sequences were represented by feature vectors of consecutive amino acid triplets and the data set was for only HPV PPIs.²⁹ The average sensitivity was 77.8%, the average specificity was 85.4% and the average accuracy of 81.6%.

A vector representation of 3 major features of the sequences was used to train other SVM models.³⁰ The features consisted of the frequency difference of amino acid triplets, relative frequency of amino acid triplets, and amino acid composition. The models were trained with different combinations of the aforementioned features with the performance of 99.4% for sensitivity, 99.6% for specificity, 99.5% for accuracy, and an MCC of 0.989. Our PPI model developed which is a deep learning-based, performed better than some of these SVM models.

The ability of the model to predict proteins of other HPV types and host PPIs was also tested. This was achieved by testing the model's performance on a data set of HPV18 and host PPIs, and a recall of about 70% was achieved; this is similar to the model's performance on an evaluation data set of HPV31 and host PPIs (recall of 76%), suggesting that the model can be

applied to predict other HPV types and host protein interaction effectively.

Finally, the model was then used to predict the possible interaction between E6 and E7 proteins of HPV18 and 31 and AKT, IQGAP1, and MMP16. From the results of the prediction, HPV E6 was found to only interact with AKT, while HPV E7 interacted with AKT, IQGAP1, and MMP16. The HPV E6 and E7 have been reported to initiate cancer by deregulating P53 and retinoblastoma protein (pRb), leading to unrestrained cell proliferation.³¹ In addition, HPV E6/E7 expression has been reported to activate the PI3K signaling pathway and contribute to the amplification or mutation of the major components in this pathway.³² The findings of the prediction suggest that HPV E6/E7 directly interact with some components of the PI3K pathway (AKT, IQGAP1, and MMP16), and could explain the increased mutations and amplifications observed in components of the PI3K pathway in HPV-positive cancers. The results also suggest that E7 may be more involved in the deregulation of the PI3K pathway because of its ability to interact with more components of the pathway. Therefore, therapies targeting the disruption of E6/E7 and AKT, IQGAP1, and MMP16 interaction in HPV-associated cancers may be helpful in reducing the severity of these cancers. The deep learning model developed for prediction is primarily computational, and the predicted interactions must be corroborated using experimental assays.

Conclusions

We were able to develop for the first time a deep learning model for the prediction of HPV and host PPIs. The findings from the model's prediction show that HPV E7 may interact more with components from the PI3K pathway (AKT, IQGAP1, and MMP16). These interactions should be a major driver in the deregulation of the pathway while E6 may contribute to activating AKT, leading to subsequent activation of other downstream oncogenes. These computationally elucidated interactions can be confirmed experimentally to provide insight into biomolecular mechanisms.

Author Contributions

SKK, EAT, CAB, and OQ conceptualized and designed the work; SS, KA-M acquired and analyzed data or SKK, SS, and KA-M interpreted data; EAT and SS drafted the article; all authors revised it critically for important intellectual content; and all authors approved the version to be published.

Acknowledgements

The authors specially thank Quaye Lab. Members of the Virology Lab, Department of Biochemistry, Cell and Molecular Biology, University of Ghana, for supporting the project.

Availability of Data and Materials

All data generated or analyzed during this study are included in this article.

SUPPLEMENTAL MATERIAL

Supplemental material for this article is available online https://github.com/schistomodels/peptide_deep_learning/.

ORCID iD

Emmanuel A Tagoe  <https://orcid.org/0000-0001-7179-1872>

REFERENCES

- Arbyn M, Weiderpass E, Bruni L, et al. Estimates of incidence and mortality of cervical cancer in 2018: a worldwide analysis. *Lancet Glob Health*. 2019;8:e191-e203. doi:10.1016/s2214-109x(19)30482-6
- de Martel C, Plummer M, Vignat J, Franceschi S. Worldwide burden of cancer attributable to HPV by site, country and HPV type. *Int J Cancer*. 2017;141:664-670. doi:10.1002/ijc.30716
- Smalley Rumfield C, Roller N, Pellom ST, Schlom J, Jochems C. Therapeutic vaccines for HPV-associated malignancies. *Immunotargets Ther*. 2020;9:167-200. doi:10.2147/ITT.S273327
- Evande R, Rana A, Biswas-Fiss EE, Biswas SB. Protein-DNA interactions regulate human papillomavirus DNA replication, transcription, and oncogenesis. *Int J Mol Sci*. 2023;24:8493-8493. doi:10.3390/ijms24108493
- Graham Sheila V. The human papillomavirus replication cycle, and its links to cancer progression: a comprehensive review. *Clin Sci*. 2017;131:2201-2221. doi:10.1042/cs20160786
- Bhattacharjee R, Das SS, Biswal SS, et al. Mechanistic role of HPV-associated early proteins in cervical cancer: molecular pathways and targeted therapeutic strategies. *Crit Rev Oncol Hematol*. 2022;174:103675. doi:10.1016/j.critrevonc.2022.103675
- He Y, Sun MM, Zhang GG, et al. Targeting PI3K/Akt signal transduction for cancer therapy. *Signal Transduct Target Ther*. 2021;6:1-17. doi:10.1038/s41392-021-00828-5
- Gui C, Ji M, Song Y, Wang J, Zhou Y. Functions and modulation of PKM2 activity by human papillomavirus E7 oncoprotein (Review). *Oncol Lett*. 2022;25:7. doi:10.3892/ol.2022.13593
- Kim D, Kim S, Koh H, et al. Akt/PKB promotes cancer cell invasion via increased motility and metalloproteinase production. *FASEB J*. 2001;15:1953-1962. doi:10.1096/fj.01-0198com
- Yerramilli VS, Ross AH, Lindberg SK, Scarlata S, Gericke A. IQGAP1 connects phosphoinositide signaling to cytoskeletal reorganization. *bioRxiv*. 2019;121:793-807. doi:10.1101/706465
- Miura K. An overview of current methods to confirm protein-protein interactions. *Protein Pept Lett*. 2018;25:728-733. doi:10.2174/0929866525666180821122240
- Ding Z, Kihara D. Computational identification of protein-protein interactions in model plant proteomes. *Sci Rep*. 2019;9:8740. doi:10.1038/s41598-019-45072-8
- Kewalramani N, Emili A, Crovella M. State-of-the-art computational methods to predict protein-protein interactions with high accuracy and coverage. *Proteomics*. 2023;23:e2200292. doi:10.1002/pmic.202200292
- Sarker IH. Machine learning: algorithms, real-world applications and research directions. *SN Comput Sci*. 2021;2:160. doi:10.1007/s42979-021-00592-x
- Hu X, Feng C, Ling T, Chen M. Deep learning frameworks for protein-protein interaction prediction. *Comput Struct Biotechnol J*. 2022;20:3223-3233. doi:10.1016/j.csbj.2022.06.025
- Eckhardt M, Zhang W, Gross AM, et al. Multiple routes to oncogenesis are promoted by the human papillomavirus-host protein network. *Cancer Discov*. 2018;8:1474-1489. doi:10.1158/2159-8290.CD-17-1018
- Cook H, Doncheva N, Szklarczyk D, von Mering C, Jensen L. Viruses. STRING: a virus-host protein-protein interaction database. *Viruses*. 2018;10:519. doi:10.3390/v10100519
- Bateman A, Martin MJ, Orchard S, et al. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res*. 2020;49:D480-D489. doi:10.1093/nar/gkaa1100

19. Gonzalez-Lopez F, Morales-Cordovilla JA, Villegas-Morcillo A, Gómez ÁM, Sánchez V. End-to-end prediction of protein-protein interaction based on embedding and recurrent neural networks. Paper presented at: IEEE International Conference on Bioinformatics and Biomedicine (BIBM); December 3–6, 2018; Madrid. doi:10.1109/bibm.2018.8621328
20. Asgari E, Mofrad MR. Continuous distributed representation of biological sequences for deep proteomics and genomics. *PLoS ONE*. 2015;10:e0141287. doi:10.1371/journal.pone.0141287
21. Ganapathiraju M, Weisser D, Rosenfeld R, Carbonell J, Reddy R, Klein-Seetharaman J. Comparative n-gram analysis of whole-genome protein sequences. In: Proceedings of the Second International Conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc; 2002:76–81. <https://dl.acm.org/doi/10.5555/1289189.1289259>
22. Vries JK, Liu X. Subfamily specific conservation profiles for proteins based on n-gram patterns. *BMC Bioinformatics*. 2008;9:72
23. Mao S, Sejdic E. A review of recurrent neural network-based methods in computational physiology. *IEEE Trans Neural Netw Learn Syst*. 2022;34:6983–7003. doi:10.1109/tnnls.2022.3145365
24. Yang L, Xia JF, Gui J. Prediction of protein-protein interactions from protein sequence using local descriptors. *Protein Pept Lett*. 2010;17:1085–1090. doi:10.2174/092986610791760306
25. You ZH, Zhu L, Zheng CH, Yu HJ, Deng SP, Ji Z. Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set. *BMC Bioinformatics*. 2014;15:S9. doi:10.1186/1471-2105-15-s15-s9
26. Li H, Gong XJ, Yu H, Zhou C. Deep neural network based predictions of protein interactions using primary sequences. *Molecules*. 2018;23:1923. doi:10.3390/molecules23081923
27. Wang Y, You Z, Li L, et al. Predicting protein interactions using a deep learning method-stacked sparse autoencoder combined with a probabilistic classification vector machine. *Complexity*. 2018;34:802–810. doi:10.1155/2018/4216813
28. Shen J, Zhang J, Luo X, et al. Predicting protein-protein interactions based only on sequences information. *Proc Natl Acad Sci USA*. 2007;104:4337–4341. 10.1073/pnas.0607879104
29. Cui G, Fang C, Han K. Prediction of protein-protein interactions between viruses and human by an SVM model. *BMC Bioinformatics*. 2012;13:S5. 10.1186/1471-2105-13-S7-S5
30. Kim B, Alguwaizani S, Zhou X, Huang DS, Park B, Han K. An improved method for predicting interactions between virus and human proteins. *J Bioinform Comput Biol*. 2017;15:1650024. doi:10.1142/s0219720016500244
31. Pal A, Kundu R. Human papillomavirus E6 and E7: the cervical cancer hallmarks and targets for therapy. *Front Microbiol*. 2020;10:3116. doi:10.3389/fmicb.2019.03116
32. Bossler F, Hoppe-Seyler K, Hoppe-Seyler F. PI3K/AKT/mTOR signaling regulates the virus/host cell crosstalk in HPV-positive cervical cancer cells. *Int J Mol Sci*. 2019;20:2188. doi:10.3390/ijms20092188