Using Algorithms for the Prediction of Low-risk and High-risk Human Papillomavirus (HPV) in Males

Ogbolu Melvin Omone[1]
BioTech Research Center, EKIK,
Óbuda University,
Budapest, Hungary,
ogbolu.melvin@biotech.uni-obuda.hu

Alex Ugochukwu Gbenimachor[2]
Dipartimento di Ingegneria Informatica,
Automatica e Gestionale "A. Ruberti",
Sapienza Università di Roma,
Rome, Italy,
gbenimachorugochuwkwu@gmail.com

Miklós Kozlovszky[3]
BioTech Research Center, EKIK,
Óbuda University,
Budapest, Hungary,
kozlovszky.miklos@nik.uni-obuda.hu

***Abstract*— This study focus on approximating the prevalence of Human papillomavirus (HPV) among both male staff and students of Ibrahim Badamasi Babangida University in Nigeria, and to analyze both high-risk(hr) and low-risk(lr) factors associated with the persistence of the virus in men which is the leading cause of cervical cancer in women after sexual intercourse. When an HPV-infected man is left untreated for a long period of time, it raises the opposite sex (women) at a future risk of cervical cancer development in the future after the occurrence of sexual intercourse. This paper model high-level python algorithms to categorize, quantify, classify, and predict the male participant(s) who belong to HPV high-risk or low-risk group using Neural Network Algorithm (NNA), Random Forest Algorithm (RFA), and K-Means Clustering Algorithm (KMCA).**

Keywords*— *Cervical Cancer, Human papillomavirus (HPV), HPV high-risk group, HPV low-risk group, K-Means Clustering Algorithm (KMCA), Neural Network Algorithm (NNA), Random Forest Algorithm (RFA)

I. Introduction

Recently, many studies have focused on the prevalence of Human papillomavirus (HPV) and its effects among women only, thereby neglecting the fact that men are the leading cause of this persistent virus [1]. HPVs are a non-enveloped and icosahedral double-stranded DNA [2] kind of viruses which are known to be a common Sexually Transmitted Infection (STI) from the male to the female during sexual intercourse [3], [4]. This viral infection is classified into low-risk HPV (lrHPV) and high-risk HPV (hrHPV), which are assigned with numbers called subtypes and causing different types of physiological pathosis. The high-risk HPV subtypes are HPV–16, 18, 31, 33, 35, 39, 45, 51, 52, 56, 58, 59, 68, 73 and 82, while the low-risk HPV subtypes are HPV–6, 11, 40, 42, 43, 44, 54, 61, 70, 72, 81, and 89 [5]. Many epidemiological studies have revealed that infections from low-risk HPV subtypes usually clear over a certain period of time without medical interventions while infections from high-risk HPV subtypes mostly result into cancerous diseases [5]. The resulting physiological pathosis mainly affects both the reproductive system of men and women; which is caused by high-risk HPV-infections (subtypes are mostly HPV-16 and 18 which are classified as oncogenic HPV subtypes) in women results into cancer of the cervix, cancer of the vulva, cancer of the vagina, and cancer of the anus, while in men, high-risk HPV-infections results into cancer of the anus and cancer of the penis [1], [6], [7]. However, low-risk HPV-infections (subtypes are mostly HPV-6 and 11 which are classified as non-oncogenic HPV subtypes) in men and women results into anogenital warts (i.e. benign epithelial tumors) [8], oral lesions(i.e. ulcer of the oral cavity/mouth ulcer), common warts, and plantar warts [9].

A. The Development of Human Papillomavirus (HPV) In Men

In 1996, a group of researchers revealed that due to the sexual behavior among men, the risk of cervical cancer development among women has been 5-times to 9-times on the increase in Spain. This means that men with multiple sexual partners and married men with extramarital sexual partners who are mostly infected by HPV-16, 18, 31, and 33 subtypes tend to put their wives at a higher risk of cervical cancer [10]. Hence, related studies have revealed that men are the utmost carriers of HPV infections [11] and it is likely to persist overtime in men than in women [1]. The persistence of HPV in men has resulted into both life-threatening and noncritical kind of diseases [12]. Certain life-threatening types of these diseases are the Anogenital Tract (AT) and Upper Aero-Digestive Tract (UADT) cancers and other kinds of carcinoma in-situ are minimally invasive cancers [4].

The most commonly observed pathway for HPV infection is sexual intercourse; HPV can only be transmitted during unsafe sexual practices through vaginal sex, oral sex, and anal sex [13]. Unsafe sexual practices is the act of having unprotected sexual intercourse [6] with an HPV-infected persons or person (mostly opposite sex) [14] which is a risk factor for the prevalence of HPV infection in men. However, it has been revealed that men who have sex with men (MSM) are at a higher risk of HPV-infection caused by anal sex. Although, men who are more at a higher risk are specifically associated with receptive anal sex and a multiple number of sexual partners [15]. When a man is HPV-infected with the HPV (subtypes such as 16 and 18), the persistent virus cannot directly lead to cancer. The virus is left untreated for a long period of time, develops into precursor lesions [15] and then begins to spread around the reproductive system. The development of HPV in men is likened to the development of an infected host cell-cycle. When a man is infected by HPV, the virus invades a host cell in the basal layer of epithelia (i.e. the innermost layer of the continuous sheets of the skin cells) through minor cuts and scrapes [16]. Overtime, the host cell begins to divide (this process is known as replication cycle), thereby replicating itself [17]. When the HPV DNA is replicated, it further leads to genetic mutations in the host cells that can possibly develop into cancer in future [18].

II. Materials and Methods

A. Study Population

This study includes over one hundred and fifty (150) male participants of Ibrahim Badamasi Babangida University in Nigeria, which were both students and staff of the university. All the participants who participated in the assessment were provided with a written consent that informed them about of the purpose of the study and the knowledge they would gain as a result of participating. However, the study remains

anonymous in order to respect and protect the participants privacy. Among many who volunteered to participate in the assessment, only 100 participants were qualified to be included in the study. The selection criteria were based on the following;

- Age: Participants who indicated to be between 18 – 60 years old.
- Demographic factor: Participants who are student/staff of Ibrahim Badamasi Babangida University in Nigeria.
- Record of sexual experience: Participants who indicated to have had sex in their lifetime, one/more sexual partners, and practice protected/unprotected sex.

B. Data Collection

An online questionnaire, known as Human Papillomavirus (HPV) Assessment Test (HAT) was developed with the use of Google forms as a tool for data collection which included complete anonymity of the participants data following the currently established General Data Protection Regulation (GDPR) privacy policies [19]. The online questionnaire was designed in form of an assessment with embedded scores/points which are automatically evaluated and summed up to show the level of HPV risk for each participant depending on each option selected while responding to the questions. The overall assessment was 70 points, whereby participants with ≤20 points are categorized as HPV low-risk carriers while participants with ≥21 points are categorized as HPV high-risk carriers. There were fifteen (15) questions included in the questionnaire which could provide answers about the participant's social and demographic data, safe/unsafe sexual practices, sexual intercourse activities/behavior, and knowledge, attitude, and perception (KAP) as related to HPV.

C. Data Analysis and Statistical Analysis

Data analysis was carried out using python programming algorithms and descriptive statistics. The data contains a detailed information of males who are likely to be infected with HPV based on the HPV-related risk factors contained in the questionnaire. These risk factors are most closely associated with HPV-infection, which systematically shows the level of risk according to their Age at First Sexual Intercourse (AFSI), Multiple Number of Sexual Partners (MNSP), Sexual Lifestyle and Practices (SLP), and Knowledge, Attitude and Perception/Practices (KAP) of HPV, HPV vaccine, and HPV-related infections. Males who tend to have a high-risk of HPV infection as categorized based on the risk factors are potential HPV carriers and are the cause of cervical cancer in females who get infected by them and are not treated for several years. As for data analysis, fifteen (15) questions were selected from the questionnaire, which was used to classify which participants are categorized as high-risk/low-risk HPV carriers which is in correlation with the total score of the assessment of each participants in the data. In order to achieve this classification, the Neural Network Algorithm (NNA), Random Forest Algorithm (RFA), and K-means Clustering algorithms were applied using python programming language.

1) Neural Network Algorithm (NNA)

The Neural Network Algorithm (NNA) is an algorithm that imitates the neuron in the human brain [20]. In this paper the NNA takes series of input variable from the male HPV dataset as input ($X_1, ..., X_n$) and then learns the dependent output data as output ($Y_1, ..., Y_n$) [21]. The input variables includes both low- and high-risk factors of HPV which are used to predict at random the output data (i.e. to predict participants who belongs to low/high-risk HPV group) [22]. The NNA takes in data as binaries (0 and 1). Hence, in order to learn multiclass variable in the case of the male HPV dataset with various scores (which helps to predict the output group: low/high-risk), the input results is taken from product operation [21].

2) Random Forest Algorithm (RFA)

Random Forest Algorithm (RFA) is an algorithm for supervised learning which is mainly used for classification problems/sample classification [23]. It can be applied to perform regression and classification for data analysis. RFA is known to have the attributes/characteristics related to decision trees (i.e. Information Gain, Gini Impurity, and Information Entropy); this algorithm will be applied to each of the features (i.e. risk factors contained in the data) in order to classify the participants into their classes (low/high-risk group). Hence, yields different decision trees [24], [25]. In this paper, RFA classifies the male HPV dataset by selecting which feature(s) mostly defines or contains a detailed information about all given classes.

3) K-Means Clustering Algorithm (KMCA)

Firstly, clustering is an unsupervised learning approach for classifying data; by partitioning/grouping the data into smaller clusters given the features represented in the data [26], [27]. Hence, K-Means Clustering Algorithm (KMCA) is the classification of n features/observation data and k integer clusters [28]. In this paper, the KMCA worked perfectly for features represented graphically as it is the most efficient kind of algorithm used for the male HPV dataset. Therefore, the high-risk HPV group were clearly separated from the low-risk group.

III. Results and Discussion

A. Sociodemographic Factors

TABLE I. Sociodemographic Factors

Factors	Count	Percentage
Age Group (years)		
18 – 25	23	23%
26 – 30	23	23%
31 – 40	41	41%
41 – 50	10	10%
51 – 60	3	3%
Marital Status		
Single	61	61%
Married	37	37%
Divorced	2	2%
Monthly Income (NGN)		
≤ 30,000	16	16%
30,000 – 60,000	23	23%
60,000 – 90,000	22	22%
90,000 – 120,000	13	13%
≥ 120,000	26	26%
Level of Education		
Elementary/Primary School	2	2%
High School/Secondary School	13	13%
Bachelor's Degree	54	54%
Master's Degree	24	24%
Doctorate Degree	7	7%

Using descriptive statistics, Table 1 reveals certain significant percentages for the population sociodemographic

factors. The population shows a higher number of middle-aged; thus 41(41%) of the participants indicated to be between the age of 31 – 40 years. Hence, the population contained few young-adults and very few older-adult, where 23(23%) indicated to be between 18 – 25 years, 23(23%) indicated to be between 26 – 30 years, 10(10%) indicated to be between 41 – 50 years and only 3(3%) indicated to be between 51 – 60 years.

Since the assessment was mostly taken by participants who work/study in an university (i.e. an academia environment), statistics reveals that most of the participants were singles; thus 61(61%) indicated to be singles, 37(37%) indicated to be married, and only 2(2%) indicated to be divorced.

This study shows that most of the participants do not earn enough income to care for their needs or pay for medical bills; thus, 16(16%) earned ≤#30,000, 23(23%) earned between #30,000 – #60,000, 22(22%) earned between #60,000 – #90,000, 13(13%) earned between #90,000 – #120,000, and 26(26%) earned ≥#120,000 above. Poorly, 16(16%) of the participants who indicated to earn ≤#30,000 monthly which is the same as the National Minimum Wage (NMW) for federal workers in Nigeria could be living in abject poverty. Since, statistics discovered that for an individual who resides in Nigeria, #43,200 is the estimated cost of living while for a family of four is estimated to be #137,600 [29], [30]. Our study shows that the participants of this assessment do not earn enough income to take care of their family in case of the married ones who earns ≤#30,000 – #120,000 which is not up to #137,600 which is estimated to carter for an entire family. Hence, due to lack of fund the population are likely to be HPV-infected but are not aware since the healthcare system of Nigeria is not free, and they are unable to pay for checkups, medications and medical bills.

The population under investigation were mostly academician, hence, statistics reveals a high level of education among the population. There were only 2(2%) who indicated to hold an elementary/primary school certificate, 13(13%) indicated to hold a high school/secondary school certificate, 54(54%) indicated to hold or currently undergoing a bachelor's degree, 24(24%) indicated to hold or currently undergoing a master's degree, and only 7(7%) indicated to hold a doctorate degree. Therefore, those who indicated to hold an elementary/primary and a high school/secondary school certificate are likely to be non-teaching staff who work in the university, 54(54%) who indicated to hold or currently undergoing a bachelor's degree and 24(24%) who indicated to hold or currently undergoing a master's degree are students of the university, and 7(7%) who indicated to hold a doctorate degree are likely lecturers/instructors of the university.

B. Data Training and Loss Function

1) Training and Validation Loss for Neural Network Algorithm (NNA)

The loss graph shows a learning curve in the Neural Network Algorithm (NNA). It computes the difference between the predicted label (for scores) and the actual label (HPV scores for entire population). The HPV Assessment Test (HAT) computes a score for each participant which can be used to predict if a participant belongs to the high-risk/low-risk group. Using the NNA, if the HPV scores for entire population computes to zero (0) [31], it means that the model prediction is perfect. In this study, where the epoch=150, computation=0 for the model [32] to predict the labels given the HPV features (risk factors contained in the data). The effect of the loss function is to make sure that each datapoint of the features are classified properly/perfectly.

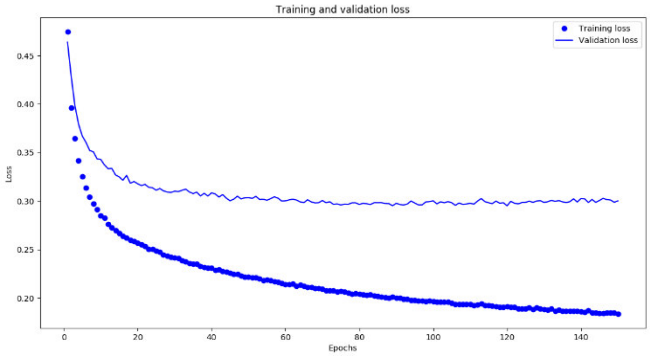

Fig 1: Loss Graph for Neural Network Algorithm (NNA)

2) Training and Validation Accuracy for Neural Network Algorithm (NNA)

The accuracy graph shows how well the Neural Network Algorithm (NNA) performed in classifying the scores (labels) given the number of input data. It shows the ratio of prediction correctness/efficacy [33] of the male HPV dataset. The Accuracy for validation and training is given as 95% and 96%, which shows that the NNA performs well in classifying each feature provided with the scores.

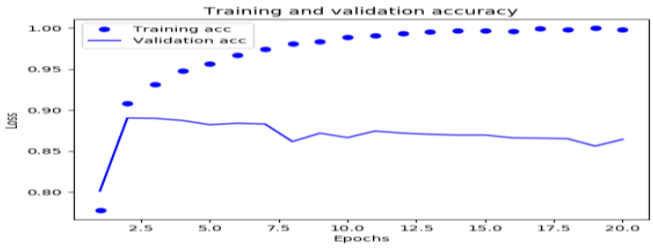

Fig 2: Accuracy Graph for Neural Network Algorithm (NNA)

3) Training and Validation Loss for Random Forest Algorithm (RFA)

In computing the loss graph for Random Forest Algorithm (RFA), the Out of the Bag (OOB) estimate was used to compute the average loss for each computed prediction from each boosted decision tree that do not contain in their respective bootstrap samples. The training of the male HPV dataset demonstrates how the OOB estimate can be measured at the addition of each new tree during the training [34].

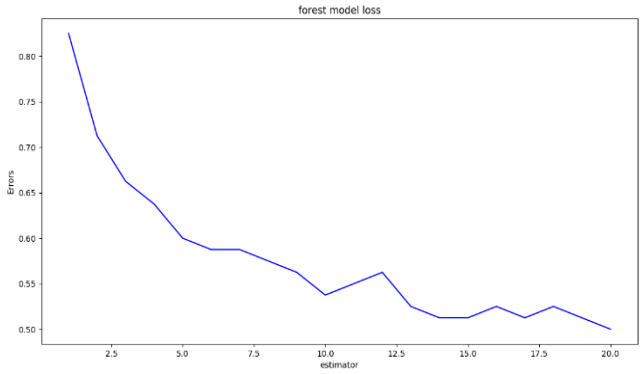

Fig 3: Loss Graph for Random Forest Algorithm (RFA)

4) Training and Validation Accuracy for Random Forest Algorithm (RFA)

The accuracy achieved in training the Random Forest Algorithm (RFA) on the male HPV dataset is 0.45/45% given 20 trees (was done using 20 estimates to classify the data).

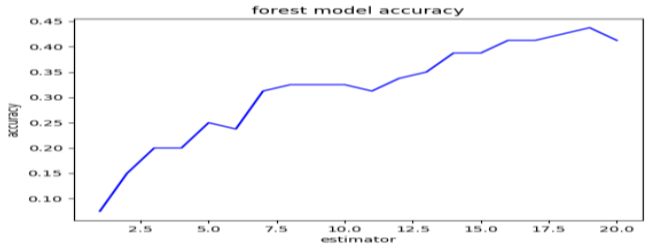

Fig 4: Accuracy Graph for Random Forest Algorithm (RFA)

5) Training and Validation Loss for K-Means Clustering Algorithm (KMCA)

In order to compute the loss function for K-Means Clustering Algorithm (KMCA) which basically tells the difference between the predicted labels and actual labels. The KMCA uses the 'elbow method' [35] which measures/computes the number of clusters in the male HPV datasets. The 'elbow method' performs a series of *k* iterations and record the Sum Square Errors (SSE) [36]. During the training of the male HPV dataset, there is a convergence from *n* clusters from *k-5* to *k-20*.

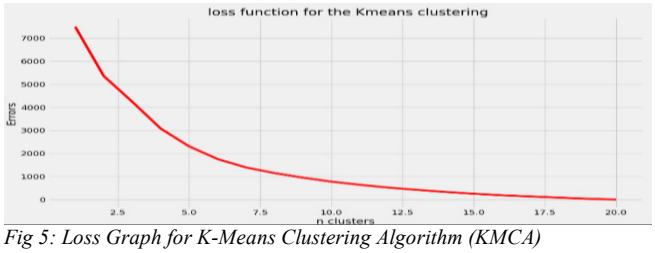

Fig 5: Loss Graph for K-Means Clustering Algorithm (KMCA)

6) Training and Validation Accuracy for K-Means Clustering Algorithm (KMCA)

The K-Means Clustering Algorithm (KMCA) gave an accuracy of 0.26 or 26% and the resultant effect on the algorithm being an unsupervised algorithm is to learn the similarities in features and group in their label clusters [28]. The result here showed that the KMCA performed poorly in classifying the features to their respective labels.

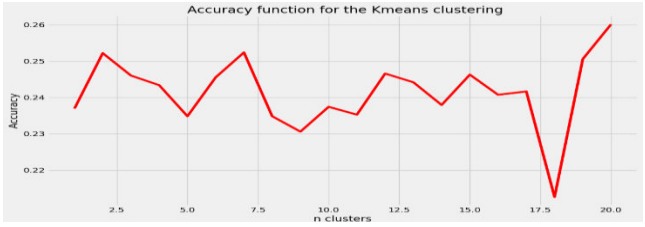

Fig 6: Accuracy Graph for K-Means Clustering Algorithm (KMCA)

C. Prediction for HPV High-risk/Low-risk Group

For classification, low-risk HPV (lrHPV) group/carriers are participants with ≤20 points while high-risk HPV (hrHPV) group/carriers are participants with ≥21 points.

1) Classification with NNA for MNSP and AFSI

As shown in Figure 7 below, NNA was used for the classification of the risk factors/features for HPV (i.e. Multiple Number of Sexual Partners (MNSP) and Age at First Sexual Intercourse (AFSI)). These features are classified into categories which are then identified by different colors given by each score labels. As mentioned above, lrHPV group are participants with ≤20 points while hrHPV group are participants with ≥21 points. The high-risk group have a higher tendency of infecting women with HPV; they are clearly identified by colors and shows participants with MNSP who started having sex at a young age (AFSI), scores are 20/70 – 25/70 points.

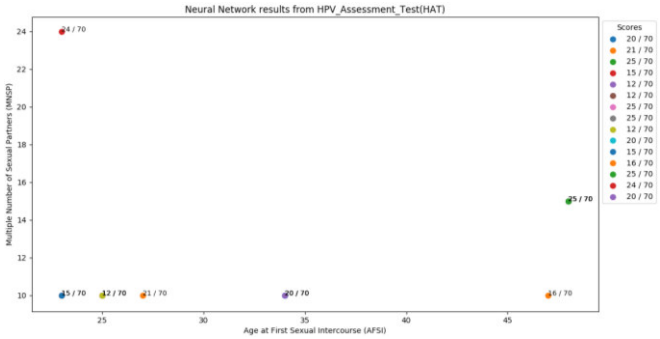

Fig 7: Classification with NNA for MNSP and AFSI

2) Clustering with KMCA for MNSP and AFSI

Using KMCA as shown in Figure 8 below, predicted labels reveals clustered features identified with each colored label. These colors are chosen at random to identify each of the labels. This tells which group given their labels have a higher tendency of infecting women with HPV and which group does not. The predicted labels are represented in the clusters of their own colors. The graph below shows a group of participants MNSP and AFSI; these features are grouped according to their scores and are described in the legend of the graph. Scores are 10/70 – 60/70 points. Each of the scores measures the risk each male-participant poses to their partners (women). The graph shows a good prediction with true labels.

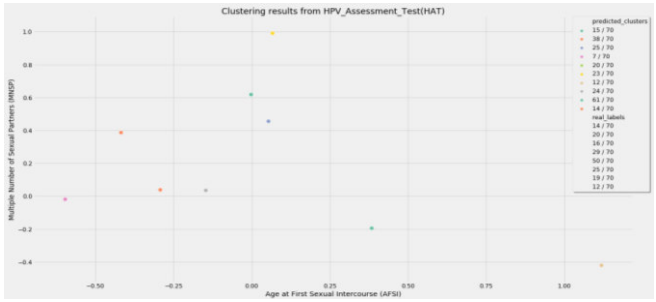

Fig 8: Clustering with KMCA for MNSP and AFSI

3) Clustering with KMCA HPV Vaccinated/Non-vaccinated Participants and HPV Infection History

Figure 9 shows the features for participants who have been vaccinated or not vaccinated against with HPV infection and those with medical history for HPV infection. Each predicted label shows the classification for each predicted cluster in the graph; it shows the cluster of features 7/70 – 61/70 point. Showing. The predicted hrHPV group in the graph are 22, 36, 38, and 61 points.

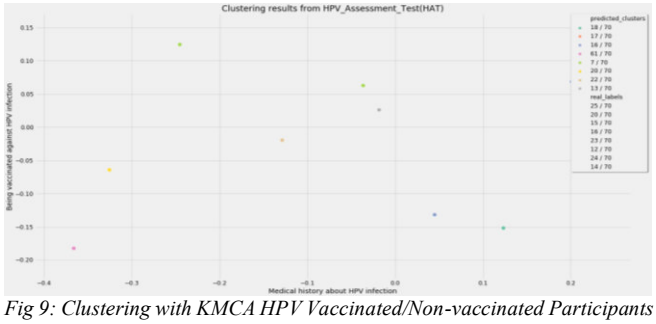

Fig 9: Clustering with KMCA HPV Vaccinated/Non-vaccinated Participants and HPV Infection History

4) Clustering with KMCA for Knowledge about HPV Genotypes and HPV Vaccinated/Non-vaccinated Participants

Using KMCA, Figure 10 shows that the participants who have no knowledge about genotypes and who have not been HPV vaccinated are at high-risk of HPV infection. The predicted clusters are used to classify and differentiate the low-risk group from the high-risk group. The data points are classified according to the predicted clusters. It also shows prediction for true labels and real labels. The legend shows predicted clusters are the high-risk HPV group.

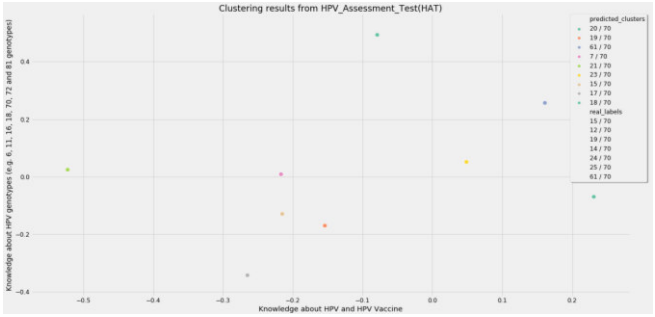

Fig 10: Clustering with KMCA for Knowledge about HPV Genotypes and HPV Vaccinated/Non-vaccinated Participants

5) Classification with NNA for SLP

Sexual Lifestyle and Practices (SLP) is used to check for the Sexual Lifestyle and Practices (oral sex, anal sex, vaginal sex, and protected/unprotected sex practices) of the participants. The graph below shows the classification between SLP for oral, anal, and vaginal sex and protected/unprotected sex practices using NNA. The male HPV dataset was well classified according to each label given the features set data points. However, some labels overlapped into single data point/features. High-risk participants have 21, 24, 29, and 50 points.

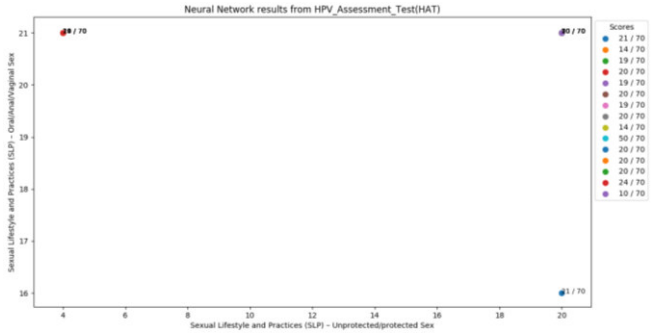

Fig 11: Classification with NNA for SLP

6) Classification with NNA for Knowledge about HPV Genotypes and HPV Vaccinated/Non-vaccinated Participants

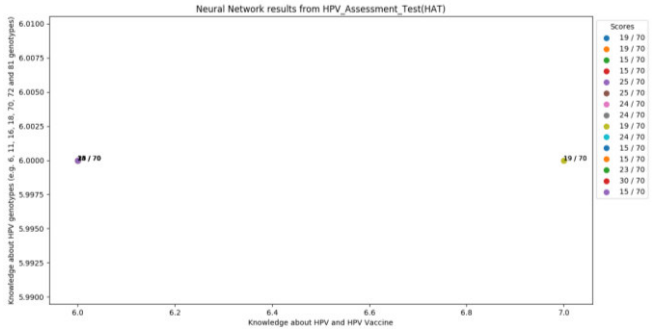

Fig 12: Classification with NNA for Knowledge about HPV Genotypes and HPV Vaccinated/Non-vaccinated Participants

Figure 12 above classifies participants who have no knowledge about HPV genotypes and who have not been vaccinated against HPV infection using data points. These data points are features sets and classified given the scores/labels. It classified participants who scored 15/70, 25/70, and 19/70 as those with high-risk tendencies of infecting women with HPV. However, it is obvious that NNA doesn't work perfectly for classifying these risk factors.

7) Classification with RFA for Knowledge about HPV Genotypes and HPV Vaccinated/Non-vaccinated Participants

As shown in Figure 13, the RFA was able to make same classification as NNA. It classified the data point features to labels as in the scores/labels for participants who have no knowledge about HPV genotype and those who have not been vaccinated against HPV infection.

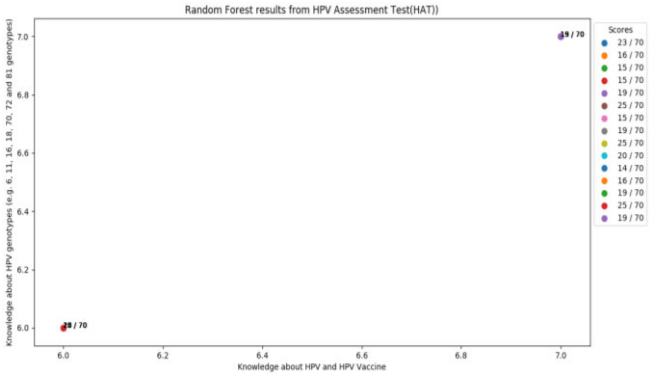

Fig 13: Classification with RFA for Knowledge about HPV Genotypes and HPV Vaccinated/Non-vaccinated Participants

8) Classification with RFA for AFSI and HPV Vaccinated/Non-vaccinated Participants

Figure 14 shows that participants who falls under the category of early Age at First Sexual Intercourse (AFSI) and those who have not been vaccinated against HPV infection belong to the HPV high-risk group. The classification for each data features into the given scores/labels are identified by different colors.

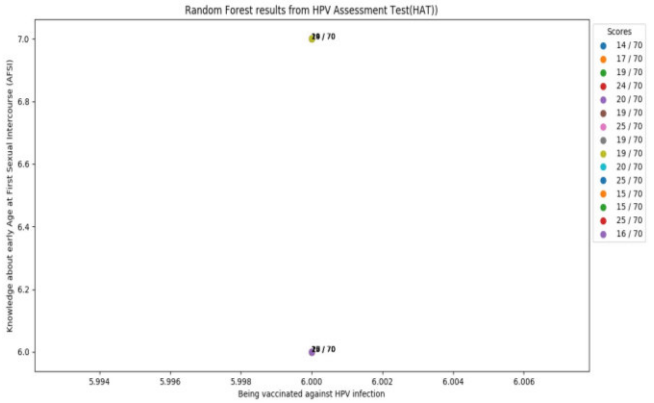

Fig 14: Classification with RFA for AFSI and HPV Vaccinated/Non-vaccinated Participants

9) Classification with RFA for SLP

Figure 15 shows result for participants with Sexual Lifestyle and Practices (oral sex, anal sex, vaginal sex, and protected/unprotected sex practices). The result shows data points with features of participants with risk given different scores/labels. Hence, the legend as shown in the figure below is the same as the classification in the NNA as in Figure 11.

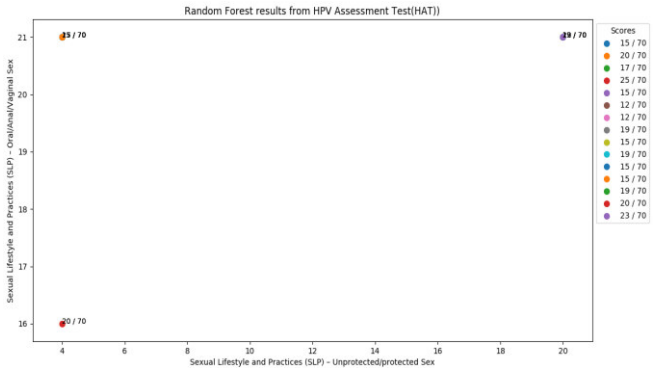

Figure 15: Classification with RFA for SLP

IV. Conclusion

The algorithms used in this paper shows good predictions with Figures 7, 8, 9, 10, 11, and 15 while Figures 12, 13, and 14 did not show good predictions. The limitation experienced in using Random Forest Algorithm (RFA) as shown in Figures 12, 13, and 14 is due to the amount of dataset used for analysis. However, Neural Network Algorithm (NNA) out-performed the K-Means Clustering Algorithm (KMCA) and the Random Forest Algorithm (RFA). Hence, it is believed that other python algorithms could do better if implemented and with the availability of a large-scale data to analyze or learn the features contained in the dataset. The purpose of this project was achieved, the algorithms classified the scores/labels given the features in the male HPV dataset that was learned during the training and validation process. In future, to get better accuracy for a better classification of the features, the Human Papillomavirus (HPV) Assessment Test (HAT) will be distributed to a larger population in order to collect large-scale data.

Acknowledgment

The authors thank the EFOP-3.6.1-16-2016-00010 project and the GINOP-2.2.1-15-2017-00073 named "Telemedicina alapú ellátási formák fenntartható megvalósítását támogató keretrendszer kialakítása és tesztelése, Hungary for their financial support.

References

[1] A. R. Giuliano, G. Anic, and A. G. Nyitray, "Epidemiology and pathology of HPV disease in males," *Gynecol. Oncol.*, vol. 117, no. 2 Suppl, pp. S15-19, May 2010.

[2] "Human Papillomavirus (HPV) Infection - Human Papillomaviruses - NCBI Bookshelf." https://www.ncbi.nlm.nih.gov/books/NBK321770/ (accessed Aug. 12, 2020).

[3] F. Cutts *et al.*, "Human papillomavirus and HPV vaccines: a review," *Bull World Health Organ*, vol. 85, no. 9, pp. 719–726, Sep. 2007.

[4] A. Castillo, "Human Papillomavirus and Carcinogenesis in the Upper Aero-Digestive Tract," in *Carcinogenesis*, K. Tonissen, Ed. InTech, 2013.

[5] O. M. Omone and M. Kozlovszky, "HPV and Cervical Cancer Screening Awareness: A Case-control Study in Nigeria," in *2020 IEEE 24th International Conference on Intelligent Engineering Systems (INES)*, Jul. 2020, pp. 145–152.

[6] M. Rodríguez-Álvarez, J. Gómez-Urquiza, H. Husein-El Ahmed, L. Albendín-García, J. Gómez-Salgado, and G. Cañadas-De la Fuente, "Prevalence and Risk Factors of Human Papillomavirus in Male Patients: A Systematic Review and Meta-Analysis," *IJERPH*, vol. 15, no. 10, p. 2210, Oct. 2018.

[7] H. Husein-El Ahmed and G. A. Cañadas-De la Fuente, "IMAGES IN CLINICAL MEDICINE. Squamous-Cell Carcinoma of the Penis with Human Papillomavirus," *N. Engl. J. Med.*, vol. 374, no. 2, p. 164, Jan. 2016.

[8] G. Ozaydin-Yavuz, S. G. Bilgili, H. Guducuoglu, I. H. Yavuz, S. Elibuyuk-Aksac, and A. S. Karadag, "Determinants of high-risk human papillomavirus infection in anogenital warts," *Postepy Dermatol Alergol*, vol. 36, no. 1, pp. 76–81, Feb. 2019.

[9] S. N. Adebamowo, O. Olawande, A. Famooto, E. O. Dareng, R. Offiong, and C. A. Adebamowo, "Persistent Low-Risk and High-Risk Human Papillomavirus Infections of the Uterine Cervix in HIV-Negative and HIV-Positive Women," *Front Public Health*, vol. 5, Jul. 2017.

[10] F. X. Bosch *et al.*, "Male sexual behavior and human papillomavirus DNA: key risk factors for cervical cancer in Spain," *J. Natl. Cancer Inst.*, vol. 88, no. 15, pp. 1060–1067, Aug. 1996.

[11] E. I. Svare, "Risk factors for genital HPV DNA in men resemble those found in women: a study of male attendees at a Danish STD clinic," *Sexually Transmitted Infections*, vol. 78, no. 3, pp. 215–218, Jun. 2002.

[12] H. A. Cubie, "Diseases associated with human papillomavirus infection," *Virology*, vol. 445, no. 1–2, pp. 21–34, Oct. 2013.

[13] N. Osazuwa-Peters *et al.*, "Understanding of risk factors for the human papillomavirus (HPV) infection based on gender and race," *Sci Rep*, vol. 9, no. 1, p. 297, Dec. 2019.

[14] Zitkute V and Bumbuliene Z: Risk Factors Affecting HPV Infection, Persistence and Lesion Progression in Women and Men, Clin Res Infect Dis 3(2): 1026, 2016.

[15] Marra, E.: Anal HPV Infection and Disease: Common and preventable, but hard to treat, 2018.

[16] A. Castillo,"HPV infection and carcinogenesis in the upper aero-digestive tract,"*Colombia Médica*, vol.42, no.2, pp.233–242, Jun. 2011.

[17] S. V. Graham, "The human papillomavirus replication cycle, and its links to cancer progression: a comprehensive review," *Clinical Science*, vol. 131, no. 17, pp. 2201–2221, Sep. 2017.

[18] "New Clues About How HPV Spreads In the Body." https://www.cancer.org/latest-news/new-clues-about-how-hpv-spreads-in-the-body.html#for_researchers (accessed Aug. 12, 2020).

[19] "General Data Protection Regulation (GDPR) – Official Legal Text." https://gdpr-info.eu/ (accessed Mar. 02, 2020).

[20] E. Lin and S.-J. Tsai, "Machine Learning in Neural Networks," in *Frontiers in Psychiatry*, vol. 1192, Y.-K. Kim, Ed. Singapore: Springer Singapore, 2019, pp. 127–137.

[21] J. Patel and R. Goyal, "Applications of Artificial Neural Networks in Medical Science," *CCP*, vol. 2, no. 3, pp. 217–226, Sep. 2007.

[22] J. A. M. Sidey-Gibbons and C. J. Sidey-Gibbons, "Machine learning in medicine: a practical introduction," *BMC Med Res Methodol*, vol. 19, no. 1, p. 64, Dec. 2019.

[23] D. Denisko and M. M. Hoffman, "Classification and interaction in random forests," *Proc Natl Acad Sci U S A*, vol. 115, no. 8, pp. 1690–1692, Feb. 2018.

[24] D. P. Bhukya and S. Ramachandram, "Decision Tree Induction: An Approach for Data Classification Using AVL-Tree," *IJCEE*, pp. 660–665, 2010.

[25] C. Kingsford and S. L. Salzberg, "What are decision trees?," *Nat Biotechnol*, vol. 26, no. 9, pp. 1011–1013, Sep. 2008.

[26] M. Z. Rodriguez *et al.*, "Clustering algorithms: A comparative approach," *PLoS ONE*, vol. 14, no. 1, p. e0210236, 2019.

[27] M. R. David and S. Samuel, "Clustering of PubMed abstracts using nearer terms of the domain," *Bioinformation*, vol. 8, no. 1, pp. 20–25, Jan. 2012.

[28] H. Alashwal, M. El Halaby, J. J. Crouse, A. Abdalla, and A. A. Moustafa, "The Application of Unsupervised Clustering Methods to Alzheimer's Disease," *Front Comput Neurosci*, vol. 13, May 2019.

[29] "Nigeria National Minimum Wage | 2018-2020 Data | 2021-2022 Forecast | Historical." https://tradingeconomics.com/nigeria/minimum-wages (accessed Jul. 02, 2020).

[30] "• Nigeria: minimum wage 2018-2020 | Statista." https://www.statista.com/statistics/1119133/monthly-minimum-wage-in-nigeria/ (accessed Sep. 05, 2020).

[31] R. Yamashita, M. Nishio, R. K. G. Do, and K. Togashi, "Convolutional neural networks: an overview and application in radiology," *Insights Imaging*, vol. 9, no. 4, pp. 611–629, Jun. 2018.

[32] A. R. Asif *et al.*, "Performance Evaluation of Convolutional Neural Network for Hand Gesture Recognition Using EMG," *Sensors (Basel)*, vol. 20, no. 6, Mar. 2020.

[33] L. Bertolaccini, P. Solli, A. Pardolesi, and A. Pasini, "An overview of the use of artificial neural networks in lung cancer research," *J Thorac Dis*, vol. 9, no. 4, pp. 924–931, Apr. 2017.

[34] S. Janitza and R. Hornung, "On the overestimation of random forest's out-of-bag error," *PLoS One*, vol. 13, no. 8, Aug. 2018.

[35] E. Demidenko, "The next-generation K-means algorithm," *Stat Anal Data Min*, vol. 11, no. 4, pp. 153–166, Aug. 2018.

[36] T. Weißer, T. Saßmannshausen, D. Ohrndorf, P. Burggräf, and J. Wagner, "A clustering approach for topic filtering within systematic literature reviews," *MethodsX*, vol. 7, Feb. 2020.