Big Data Approach for Epidemiology and Prevention of HIV/AIDS

[1] Mrs. R. Indra, [2] R. Pavithra

[1]Associate Professor, [2]Research Scholar
[1,2]Department of Computer Science
[1,2]Shrimati Indira Gandhi College, Tiruchirappalli, Tamil Nadu, India-620002

Abstract : Nowadays, the disease is spreading and becoming noxious to the society inattentive of hospitalization that is present. Toxic diseases are the disorder by organisms, such as bacteria, viruses, fungi, or parasites, which happened in a normal body. Some toxic syndromes pass from one individual to another individual, some are transferred due to animal's bite or insects, and others may happen by consuming contaminated water or food or by getting exposed to the organisms which are present in the environment. AIDS becomes a rapidly spreading and turning the life to death disease. HIV spreads from one individual to another individual in the population, in many different ways that may be due to semen and blood. The study of disease is called pathology, which includes the study of cause. This research work mainly focuses on the prediction of disease like HIV/AIDS using supervised learning system.

***IndexTerms* - Machine Learning, HIV, Supervised Learning, Unsupervised Learning, Classification, Data Mining.**

I. INTRODUCTION

The mining sector is the industry most affected by the epidemics of HIV/AIDS and tuberculosis. The existence of very high risks HIV/AIDS and tuberculosis in the mining industry has been acknowledged since the 1980s, but little is being done to address the situation. They state that the current policies and regulations in the mining industry have had little effect in the prevention and control of the epidemics of HATS [1][2]. The other challenge is that the prevention and control of each disease though necessary, is insufficient in itself because of the interrelationship of the diseases. Many researchers state that many barriers exist in the way that tuberculosis and HIV/AIDS are perceived [3][4]. They challenge that failure to successfully integrate HATS control programs has the potential to threaten the viability of both programmes and note that a coordinated approach is needed if the increased incidence of HATS is to be reversed globally in the mining industry. The Millennium Development Goals acknowledge the global challenge of tuberculosis and its association with the HIV/AIDS pandemic. They also highlight the need of a coordinated approach supported at the highest level. In addition, they note that a single strategy has been seen to be limited to succeed in addressing the complex challenges of HATS epidemic [5][6]. In Africa TB control remains weak and gains can still be made through strengthening of basic disease controls. It is note that HIV/AIDS and Silicosis have severe effects on tuberculosis risk to mine workers due to their multiplicative interactions. The prevalence of silicosis in the mines leads to more mine workers being affected by tuberculosis. The mining industry has a higher risk of tuberculosis due to factors such as silica dust, crowded living conditions and HIV/AIDS infection [7][8]. In developing countries, historically the tuberculosis incidence rates in the gold mines have been always higher than the national average. The higher TB incidence rates in the mines have been attributed to higher rates of exposure to silica dust and silicosis (silicosis increases risk of tuberculosis by up to 3 times), the HIV/AIDS epidemic (HIV/AIDS increases risk of tuberculosis by up to ten times) and the environmental factors associated with the mines[9][10].

II. RELATED WORKS

Oliveira, Alexandra, et al [11] This paper aims to identify the main factors influencing reporting delays of HIV-AIDS cases within the portuguese surveillance system. The used methodologies included multilayer artificial neural networks (MLP), naive bayesian classifiers (NB), support vector machines (SVM) and the k-nearest neighbor algorithm (KNN). The highest classification accuracy, precision and recall were obtained for MLP and the results suggested homogeneous administrative and clinical practices within the reporting process. Guidelines for reductions of the delays should therefore be developed nationwise and transversally to all stakeholders.

Das, Nivedita, Manjusha Pandey, and Siddharth Swarup Rautaray [12] The HIV/AIDS is actual of an occurrence disease to proceed with to be an important global challenge of health, and actual, to bear HIV-1 virus burden testing is more and more needed at the point of care (POC). The presence of Big Data is everywhere. It is not actually data, but is a concept which actually explains about the gathering of data, organizing the data, analyzing the data and getting information out of the data. More applications are created everyday to extract the value from it which is professional and practical. The use of Big Data technologies in enterprise data warehouse and business intelligence results in better business insights and decisions. Now, Big Data analytics recently used in the point-of-care delivery and disease penetration. Big Data analytics tools are essential and useful tools, which gives strength to companies to analyze entire data related to their customers and the flea market in which they perform. As this data holds a large amount of information concerning the specific type, commodity, client service satisfaction, and client sentiment, many companies have taken the use of Big Data analytics tool.

Bisaso, Kuteesa R., et al [13] The researchers survey published studies that make use of ML techniques in HIV clinical research and care. An advanced search relevant to the use of ML in HIV research was conducted in the PubMed biomedical database. The survey outcomes of interest include data sources, ML techniques, ML tasks and ML application paradigms. A growing trend in application of ML in HIV research was observed. The application paradigm has diversified to include practical clinical application, but statistical analysis remains the most dominant application. There is an increase in the use of genomic sources of data and high performance non-parametric ML methods with a focus on combating resistance to antiretroviral therapy (ART). There is need for improvement in collection of health records data and increased training in ML so as to translate ML research into clinical application in HIV management.

Das, Nivedita, et al [14] It is projected to develop a centralized patient monitoring system using big data. Health care is the conservation or advancement of health along the avoidance, interpretation and medical care of disorder, bad health, abuse, and other substantial and spiritual deterioration in mortal. One of the best areas where big data can be used to form an advance medical care. Medical care has the potential to reduce costs of treatment, predict outbreaks of epidemics, avoid preventable diseases and improve the quality of life in general. Average human lifespan is increasing along world population, which poses new challenges to today's treatment delivery methods. . A disease is a particular abnormal condition that affects part or all of an organism not caused by external force and that consists of a disorder of a structure or function, usually serving as an evolutionary disadvantage. The study of disease is called pathology, which includes the study of cause. This paper mainly focuses on the prediction of disease like HIV/AIDS using R programming.

Zhang, Qingpeng, et al [15] In this research, the researchers examined the feasibility of using search query data to predict the number of new HIV diagnoses in China. We identified a set of search queries that are associated with new HIV diagnoses in China. The researchers developed statistical models (negative binomial generalised linear model and its Bayesian variants) to estimate the number of new HIV diagnoses by using data of search queries (Baidu) and official statistics (for the entire country and for Guangdong province) for 7 years (2010 to 2016).

Search query data were positively associated with the number of new HIV diagnoses in China and in Guangdong province. Experiments demonstrated that incorporating search query data could improve the prediction performance in nowcasting and forecasting tasks.

III. PROBLEM STATEMENT

In the existing system, methodologies included multilayer artificial neural networks (MLP), naive bayesian classifiers (NB), support vector machines (SVM) and the k-nearest neighbor algorithm (KNN) for the detection of HIV infection. The researchers survey published studies that make use of ML techniques in HIV clinical research and care. An advanced search relevant to the use of ML in HIV research was conducted in the PubMed biomedical database. The survey outcomes of interest include data sources, ML techniques, ML tasks and ML application paradigms.

- The reduced classification accuracy in the AIDS classification system.
- Decreased precision and increased false positive rate.
- Increased error rates like Relative Absolute Error (RAE), Root Mean Squared Error (RMSE), Root Relative Squared Error (RRSE).

IV. PROPOSED METHODOLOGY

In this proposed system, MapReduce operation is done on the HIV dataset to reduce the size of the dataset. In the mapper operation, the three feature selection techniques like Chi-Square analysis, Information Gain, Gain Ratio are used and Information Energy of these three techniques are calculated. In the reducer operation, the Information energy of the feature is considered as the Key-Pair. According the value of the Key-value pair, the important features are extracted. The result obtained in the feature selection phase is named as reduced dataset. Using supervised ML technique like ANN is applied on the reduced dataset to find out the classification of AIDS among the affected patients. The accuracy of the ANN is then compared with KNN and SVM classification Methods.

4.1 Data Collection

The HIV and AIDS Reporting Data Set is used to:

- Identify the groups at risk of HIV infection in England
- Monitor the short- and long-term clinical outcomes of people living with HIV infection
- Monitor the effectiveness of the national policies and guidance
- Adapt and refine interventions, as appropriate.

4.2 Hadoop Distributed File System

HDFS is the primary distributed storage used by Hadoop applications. A HDFS cluster primarily consists of a NameNode that manages the file system metadata and DataNodes that store the actual data. Hadoop is a software framework that supports data-intensive distributed applications under a free license. It enables applications to work with thousands of computational independent computers and Hadoop was derived from Google's MapReduce and Google File System (GFS) papers.

4.3 MapReduce

Hadoop Map-Reduce is a software framework for writing applications easily which processes vast amounts of data in parallel on large clusters of commodity hardware in a reliable, faulttolerant manner. A Map-Reduce job usually splits the input data-set into independent chunks which are processed by the map tasks in a completely parallel manner. The framework sorts the outputs of the maps, which are then input to the reduce tasks. Typically, both the input and the output of the job are stored in a file-system. The framework takes care of scheduling tasks, monitoring them and re-executes the failed tasks. Typically, the compute nodes and the storage nodes are the same, that is, the Map-Reduce framework and the Distributed File System are running on the same set of nodes [18].

4.4 Feature Selection

Features at the data set are analyzed. The analyzed data features are classified to detect the condition, of the heart to identify whether it is normal or abnormal. Also, it is aimed to extract the useful information from large volumes of dataset collected from various sources.

Feature Analysis: At the first stage, data are proposed and some features for disease detection are analyzed. In the feature analysis phase, the mean value for all the intervals is calculated.

Reducing Phase: In reducing phase, map reduce function is used to merge the values from map function into a single result. It reduces a set of intermediate values which share a key to a smaller set of values.

Mapping Phase In a mapping phase, first mapped tokenizes the document and emits an intermediate key-value pair for every record. After this process each of these elements will then be sorted by their key [16][17][19][20][21][22].

5.5 Classification Technique

Artificial Neural Network (ANN) is an efficient computing system whose central theme is borrowed from the analogy of biological neural networks. ANNs are also named as “artificial neural systems,” or “parallel distributed processing systems,” or “connectionist systems.” ANN acquires a large collection of units that are interconnected in some pattern to allow communication between the units. In order to form a feed-forward multi-layer in MLP, the collection of non-linear neurons is connected to one another. This technique is known to be very useful for prediction and classification issues. Cross-validation is used to determine the ‘optimal’ number of hidden layers and neurons which were relied on the experimental design of the AIDS classification framework. These units, also referred to as nodes or neurons, are simple processors which operate in parallel [23].

V. RESULTS AND DISCUSSION

5.1 Description of the Dataset

Table 1 depicts the description of the HIV dataset.

Table 1: Description about the Dataset

Feature index	Feature name
1	Indicator
2	Year
3	Geography
4	IPS
5	Race
6	Sex
7	Age Group
8	Misc
9	Rate
10	Cases
11	Population
12	Season in which the analysis was performed.
13	Age at the time of analysis
14	Childish diseases
15	Accident or serious trauma
16	Surgical intervention
17	High fevers in the last year
18	Frequency of alcohol consumption
19	Smoking habit
20	Number of hours spent sitting per day

5.2 Performance Analysis of the Feature Selection Method for the Classification Methods

Figure 1 depicts the graphical representation of the accuracy of the proposed feature selection method and existing methods by using classification methods like ANN, NB and SVM.

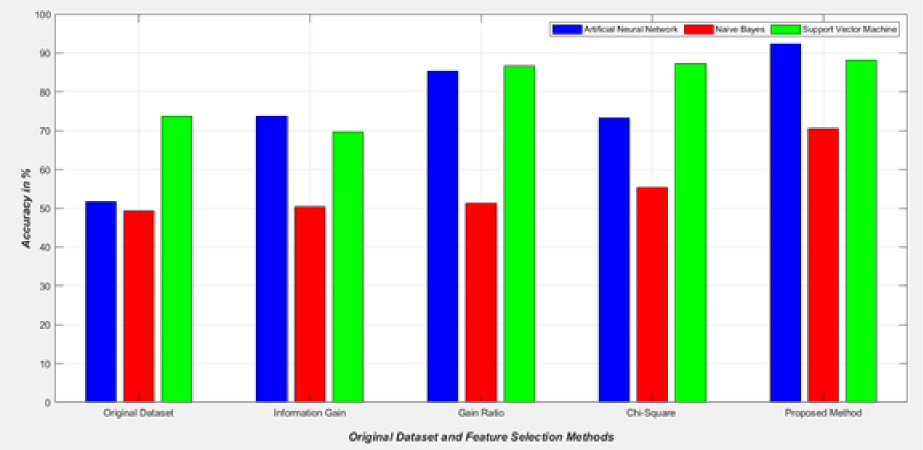

Figure 1: Graphical representation of the accuracy of the proposed feature selection method with existing feature selection methods using ANN, NB and SVM

Figure 2 depicts the graphical representation of the Kappa Statistic of the proposed feature selection method and existing methods by using classification methods like ANN, NB and SVM.

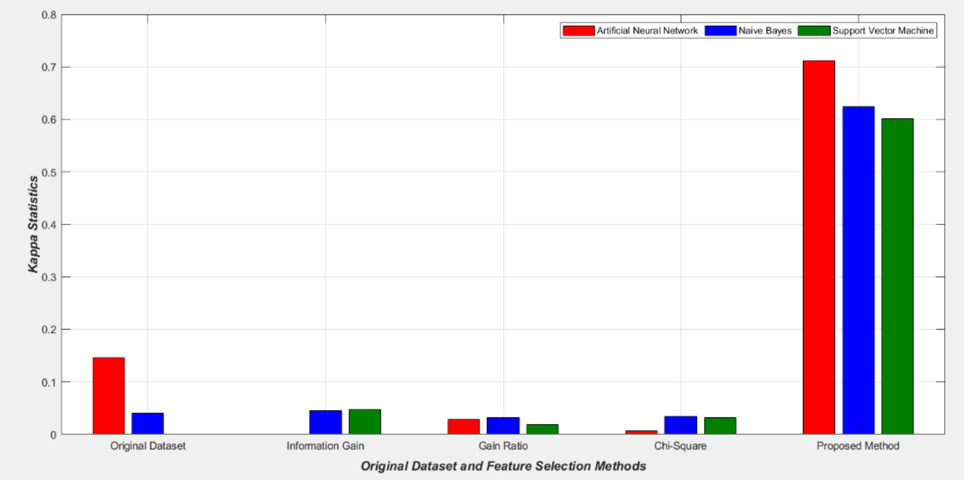

Figure 2: Graphical representation of the Kappa Statistic value of the proposed feature selection method with existing feature selection methods using ANN, NB and SVM

Figure 3 depicts the graphical representation of the TPR value of the proposed feature selection method and existing methods by using classification methods like ANN, NB and SVM.

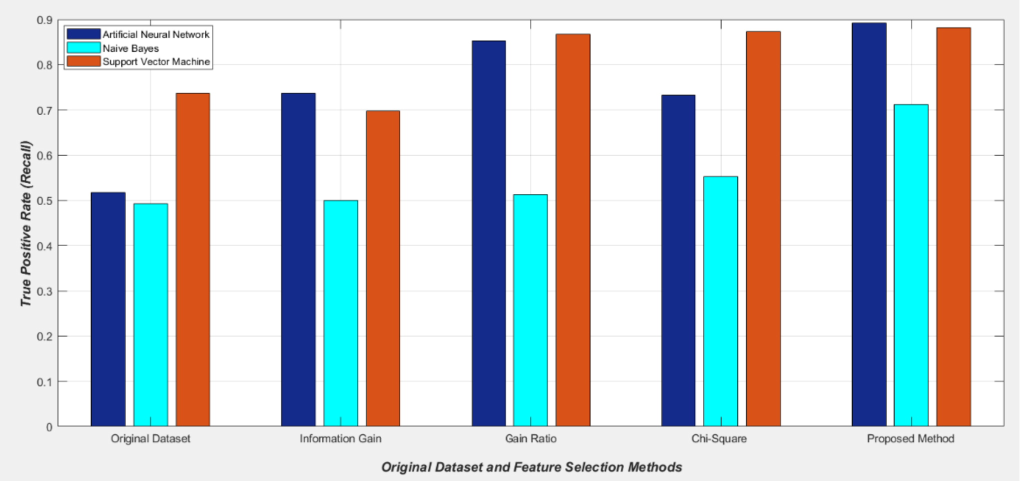

Figure 3: Graphical representation of the Kappa Statistic value of the proposed feature selection method with existing feature selection methods using ANN, NB and SVM

Figure 4 depicts the graphical representation of the FPR value of the proposed feature selection method and existing methods by using classification methods like ANN, NB and SVM.

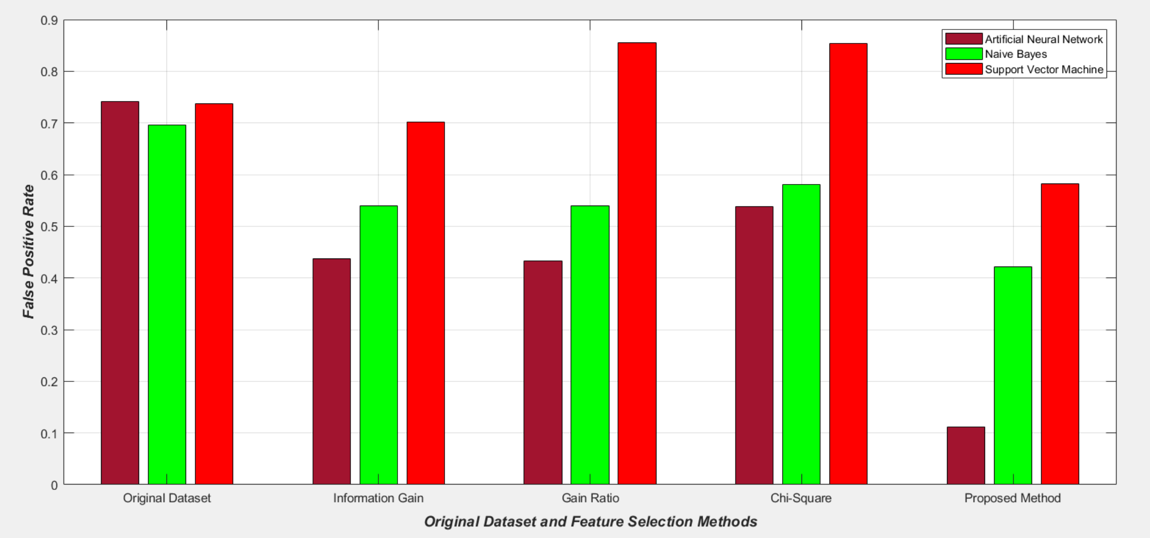

Figure 4: Graphical representation of the False Positive value of the proposed feature selection method with existing feature selection methods using ANN, NB and SVM

Figure 5 depicts the graphical representation of the Precision value of the proposed feature selection method and existing methods by using classification methods like ANN, NB and SVM.

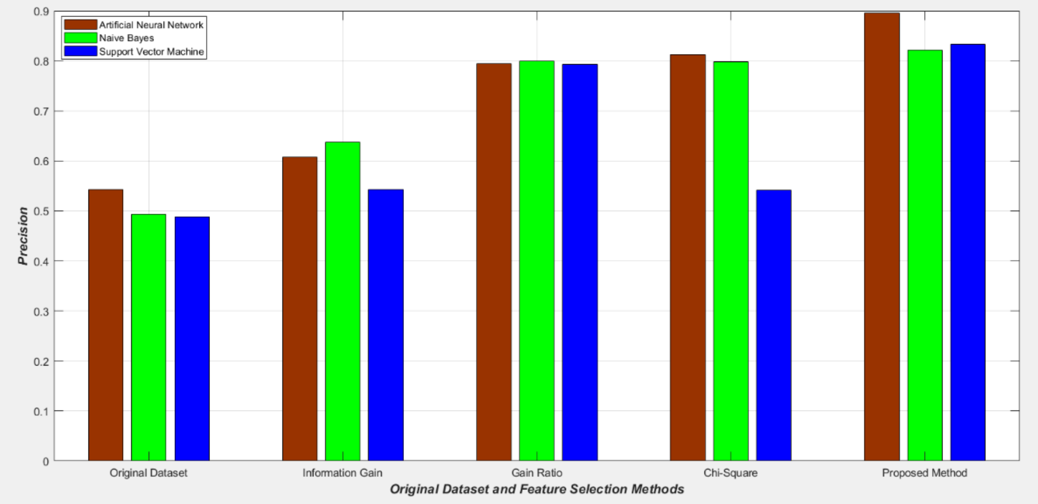

Figure 5: Graphical representation of the Precision value of the proposed feature selection method with existing feature selection methods using ANN, NB and SVM

Figure 6 depicts the graphical representation of the F-Measure value of the proposed feature selection method and existing methods by using classification methods like ANN, NB and SVM.

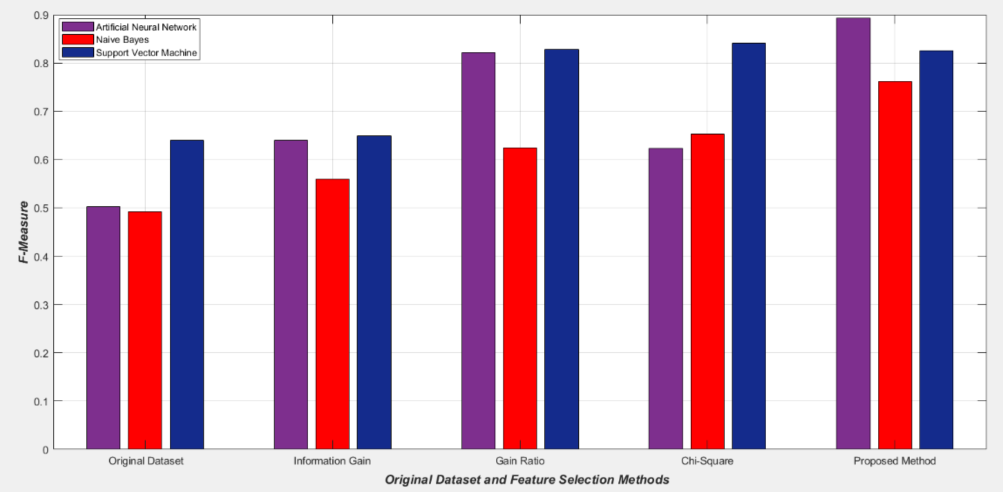

Figure 6: Graphical representation of the F-Measure value of the proposed feature selection method with existing feature selection methods using ANN, NB and SVM

Figure 7 depicts the graphical representation of the Mean Absolute Error (MAE) value of the proposed feature selection method and existing methods by using classification methods like ANN, NB and SVM.

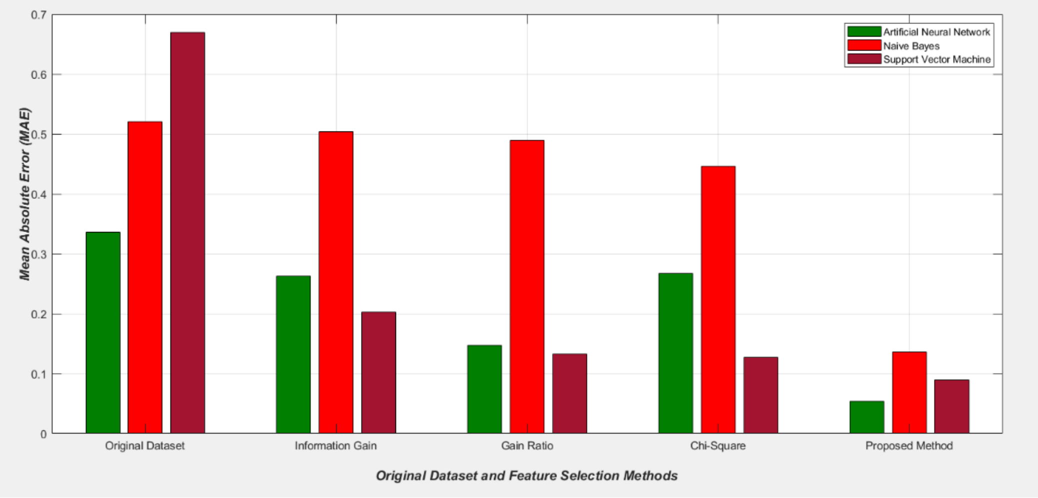

Figure 7: Graphical representation of the Mean Absolute Error (MAE) of the proposed feature selection method with existing feature selection methods using ANN, NB and SVM

Figure 8 depicts the graphical representation of the Root Mean Squared Error (RMSE) value of the proposed feature selection method and existing methods by using classification methods like ANN, NB and SVM.

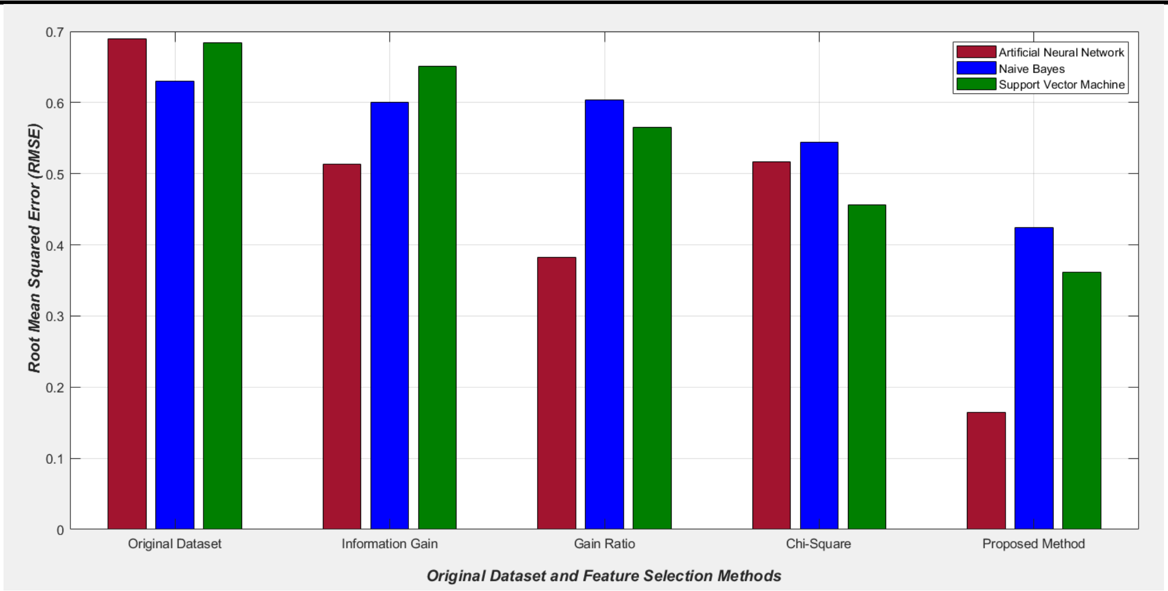

Figure 8: Graphical representation of the Root Mean Squared Error (RMSE) of the proposed feature selection method with existing feature selection methods using ANN, NB and SVM

Figure 9 depicts the graphical representation of the Relative Absolute Error (RAE) value of the proposed feature selection method and existing methods by using classification methods like ANN, NB and SVM.

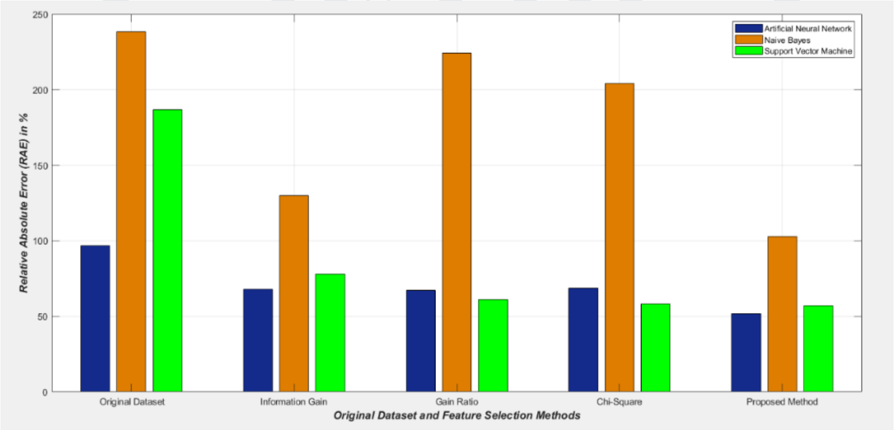

Figure 9: Graphical representation of the Relative Absolute Error (RMSE) of the proposed feature selection method with existing feature selection methods using ANN, NB and SVM.

VI. CONCLUSION

In this work, feature selector is the goal of reducing redundant and irrelevant features. The data classification has improved by picking only the most relevant features. The performance of the feature selector has estimated concerning three quality criteria such as the number of selected elements, the detection performance of classifiers, and time is taken to build the model with the HIV/AIDS dataset. The proposed system can more explicitly state as follows:

- Feature selector for choosing only most relevant features for supervised. The system has pointed at making enhancements over the present work in three aspects such as the decrease in feature set, increase in classification accuracy, and finally, reducing the running time of reaching the goal.
- The result of feature selector imparts higher classification accuracy rate for some dataset with minimum selected features and minimum running time.
- The proposed features and learning paradigm hybrid feature selector are promising strategies to be applied to any data classification problems.

REFERENCES

[1] Aavikko, M. (2014) Identification of novel tumor predisposition families and underlying genetic defects.

[2] Adrover, C., Bodnar, T., Huang, Z., Telenti, A. ,Salathé, M. (2015) Identifying adverse effects of HIV drug treatment and associated sentiments using Twitter. JMIR public health and surveillance, Vol. 1, No. 2.

[3] Alkema, L., Chou, D., Hogan, D., Zhang, S., Moller, A.-B., Gemmill, A., Fat, D.M., Boerma, T., Temmerman, M. ,Mathers, C. (2016) Global, regional, and national levels and trends in maternal mortality between 1990 and 2015, with scenariobased projections to 2030: a systematic analysis by the UN Maternal Mortality Estimation Inter-Agency Group. The Lancet, Vol. 387, No. 10017, pp. 462-74.

[4] Alkhatib, G. (2009) The biology of CCR5 and CXCR4. Current Opinion in HIV and AIDS, Vol. 4, No. 2, p. 96.

[5] Allers, K., Hütter, G., Hofmann, J., Loddenkemper, C., Rieger, K., Thiel, E. ,Schneider, T. (2011) Evidence for the cure of HIV infection by CCR5Δ32/Δ32 stem cell transplantation. Blood, Vol. 117, No. 10, pp. 2791-9.

[6] Ances, B.M. ,Ellis, R.J. (2007) Dementia and neurocognitive disorders due to HIV-1 infection, Seminars in neurology, Vol. 27, Copyright© 2007 by Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA., pp. 086-92.

[7] Antell, G.C., Dampier, W., Aiamkitsumrit, B., Nonnemacher, M.R., Jacobson, J.M., Pirrone, V., Zhong, W., Kercher, K., Passic, S. ,Williams, J.W. (2016) Utilization of HIV-1 envelope V3 to identify X4-and R5-specifi c Tat and LTR sequence signatures. Retrovirology, Vol. 13, No. 1, p. 32.

[8] Arts, E.J., Le Grice, S.F. (1997) Interaction of retroviral reverse transcriptase with template–primer duplexes during replication. Progress in nucleic acid research and molecular biology, Vol. 58, pp. 339-93.

[9] Bagri, A., Gurney, T., He, X., Zou, Y.-R., Littman, D.R., TessierLavigne, M. ,Pleasure, S.J. (2002) The chemokine SDF1 regulates migration of dentate granule cells. Development, Vol. 129, No. 18, pp. 4249-60.

[10] Bass, L.H., Washington, C.M. (2015) Infection Control in Radiation Oncology Facilities. Principles and Practice of Radiation Therapy, p. 178.

[11] Oliveira, Alexandra, et al. "Data mining in HIV-AIDS surveillance system." *Journal of medical systems* 41.4 (2017): 51.

[12] Das, Nivedita, Manjusha Pandey, and Siddharth Swarup Rautaray. "A Big Step for Prediction of HIV/AIDS with Big Data Tools." *Advances in Computer Communication and Computational Sciences*. Springer, Singapore, 2019. 37-46.

[13] Bisaso, Kuteesa R., et al. "A survey of machine learning applications in HIV clinical research and care." *Computers in biology and medicine* 91 (2017): 366-371.

[14] Das, Nivedita, et al. "Detection and Prevention of HIV AIDS Using Big Data Tool." *2018 3rd International Conference for Convergence in Technology (I2CT)*. IEEE, 2018.

[15] Zhang, Qingpeng, et al. "Using internet search data to predict new HIV diagnoses in China: a modelling study." *BMJ open*8.10 (2018): e018335.

[16] Poornappriya, T. S., and M. Durairaj. "High relevancy low redundancy vague set based feature selection method for telecom dataset." *Journal of Intelligent & Fuzzy Systems*, Preprint: 1-18.

[17] M. Durairaj, T S Poornappriya, "Choosing a spectacular Feature Selection technique for telecommunication industry using fuzzy TOPSIS MCDM.", *International Journal of Engineering & Technology*, 7 (4) (2018) 5856-5861.

[18] M. Durairaj, T. S. Poornappriya, "Importance of MapReduce for Big Data Applications: A Survey", *Asian Journal of Computer Science and Technology*, Vol.7 No.1, 2018, pp. 112-118.

[19] M. Lalli, V.Palanisamy,(2016), "Filtering Framework for Intrusion Detection Rule Schema in Mobile Ad Hoc Networks", International Journal of Control Theory and Applications –(IJCTA),9(27), pp. 195-201, ISSN: 0974-5572

[20] M. Lalli, V.Palanisamy,(2017), "Detection of Intruding Nodes in Manet Using Hybrid Feature Selection and Classification Techniques", Kasmera Journal, ISSN: 0075-5222, 45(1) (SCIE)(Impact Factor:0.071).

[21] M. Lalli, V.Palanisamy, (Sep 2014), "A Novel Intrusion Detection Model for Mobile Adhoc Networks using CP-KNN", International Journal of Computer Networks & Communications- (IJCNC), Vol.6, No.5, ISSN:0974-9322.

[22] M. Lalli, "Statistical Analysis on the KDD CUP Dataset for Detecting Intruding Nodes in MANET", *Journal of Applied Science and Computations*, Volume VI, Issue VI, JUNE/2019, 1795-1813.

[23] M. Lalli, "Intrusion Detection Rule Structure Generation Method for Mobile Ad Hoc Network", *Journal of Emerging Technologies and Innovative Research*, June 2019, Volume 6, Issue 6, 835-843.