www.revistageintec.net
ISSN: 2237-0722
revista GEINTEC
Gestão, Inovação e Tecnologias

Automatic Verbal Autopsy Classification Using Multinomial Logistic Regression Classifier by Using Recursive Feature Elimination

Zainab Mohanad Issa Ansaf[1]; Dr. Shaheda Akthar[2]

[1]Research Scholar, Department of Computer Science & Engineering, Acharya Nagarjuna University, Guntur, A.P, India.

[1]zainab.mohanad91@gmail.com

[2]Head of the Department of Computer Science, Government College for Women(A), Guntur, A.P, India.

[2]shahedaakthar76@gmail.com

Abstract

Verbal autopsy is one of the finest medical process to identify automatically the cause of a death afore medical ascendant entities will certify it. Identifying the exact cause is intricate and fuzzy in nature. The dataset with an exact cause of death is a paramount implement for every country to make the presage about the life style and medical facilities available to the people. Multinomial logistic regression was utilized in our study to relegate the exact cause of death. We used standard datasets like PHMRC and Matlab which were potentially accepted in medical field. The reason to utilize the Multinomial logistic Regression is that most of the dataset is consisting of 0 and 1 values which betoken the presence and absence of value in the attribute. We used three standard metrics like the sensitivity, Chance Corrected Concordance (CCC) and Cause-specific mortality fraction (CSMF) for a comparison of our model with precedent models like Insilico VA, Tariff and InterVA-4. Computed results show that proposed model is better than the precedent models.

Key-words: Verbal Autopsy, Cause of Death, Multinomial Logistic Regression, Chance Corrected Concordance (CCC), Cause-specific Mortality Fraction (CSMF).

1. Introduction

Identification of categorical cause of death is mostly intricate and cumbersome process. Anteriorly there is no standard technical process through which can do to extract the exact cause of death of a person. Under developing counties, where death of people conventionally transpires outside rather than hospitals. Verbal Autopsy is done through two experts Medicos where they will follow the guidelines provided by WHO (World Health Organization) and codes provided by ICD (international relegation of Diseases). These two Medicos amass the information about person death by inquiring his close relatives. If there are any controversies, subsist in their data and result then the quandary can be referred to the senior Medico to resolve. The results concerned with the manual autopsy there subsists controversies. The verbal autopsies are increasingly becoming a vital factor where the identification of exact cause of people's death can be serviceable for the countries to make health cognate policies and rules. Physician based classification of cause of death was unaccepted some times and they are costly.

Automatic Verbal Autopsy is the process of identification of exact concrete cause of by utilizing computer implements and cognate algorithms. These algorithms work on the antecedently subsisting Verbal autopsy datasets. Automatic Verbal Autopsy has received much attention in the last few years. Automatic Verbal Autopsy engendering considerable interest in terms of flexibility applying methods and preserving of time. Within the next few years Automatic Verbal Autopsy liable to become a paramount component in the medical field. There is a considerable amount of literature on Verbal Autopsy.

2. Motivation Behind this Work

[1] Established a link between precise cause of death information systems and identify the disease control priorities to availing to detect emerging epidemics. [2] proposed a design which was cost efficacious and nationally representative sample for sample vital registration. [3] in his study he made a good interpretation of Verbal Autopsy data and cognate cause of death, much fixated on the maintaining the structured, quantitative and qualitative records in terms of pristinely biomedical frame work. [4] developed an incipient method called Tariff for Automatic Verbal Autopsy and concluded that this method is transparent, intuitive and flexible and under goes rigorous testing. A new InterVA-4 model was developed, [5] which takes into account of new probabilistic model for interpreting the Verbal Autopsy data which follows the international Classification of disease version. [6] InsilicoVA was a statistical tool, which takes into account of data augmentation approach to reconcile individual cause of death with the population cause of death distribution. [7] Machine learning based Random forest algorithm was acclimated to predict the cause of death and distinguish the cause of death. [8] More computation approach was proposed in this by considering the only consequential theoretical results and empirical analysis in the data and neglecting the postulations made by Medicos, expert algorithms and parametric statistical postulations.[9] Naïve Bayes Classifier was utilized in Automatic Verbal Autopsy and performance was compared with other

Medico-predicated relegation. In [10] authors investigated validity of few models like InterVA-4, Random-Forest, Simplified symptom pattern, Traiff Method, King-Lu, Medico review of VA forms and proved that Tariff, Simplified Symptom pattern and Random-Forest performs well compared with InterVA-4. [11] In this author has analyzed and studied the working of probabilistic approach of Bayesian probability model for Automatic Verbal Autopsy. [12] In this authors has investigated the working condition of Tariff 1.0 and rigorous methods, were adopted to surmount its pitfalls in a revision of Tariff 2.0.

The aim of research is to applying the classification algorithm and withal minimizing the dimensionality of data. In this we used multinomial logistic regression (MLR) for classification and recursive feature elimination (RFE) for dimensionality truncation by culling the felicitous features. Section II describes about datasets used for computation. Section III provide framework for technical information on MLR and RFE. Section IV provides the detailed information on metrics used to access the performance of various models. Section V discussion on results of various models. Section VI includes Conclusion.

3. Data Sets Used

To analyze the performance of proposed model and previous[9] [13] models we used the Verbal Autopsy datasets from two demographic surveillance sites in Agincourt south Africa [Kahn K, Collinson], and Matlab, Bangladesh [Matlab]. Table I describes the description about the datasets and their related features. Other datasets came from Population Health Metric Consortium (PHMRC).

TABLE I - (Different Datasets used for study)

S.No	Dataset Name	No of Rows	No of Columns
1	Agincourt	5823	90
2	Matlab	2000	215
3	PHMRC_IHME_allSites_Adult_12-69yrs	4654	225
4	PH MRC_IHME_India_Adult_12-69yrs	1233	225
5	PHMRC_IHME_allSites_Child_28days-11yrs	2064	135
6	PHMRC_IHME_India_Child_28days-11yrs	948	135

From the TABLE I the dataset Agincourt consisting of 5823 rows and 90 columns with double Medico coding. Matlab dataset consisting of 2000 rows and 215 columns which was extracted

Received: 13.10.2021 – Accepted: 08.11.2021

from single coding from expertise Medico. PHMRC predicated dataset with characteristics of Adult in an age group of 12-69 years and Child in a age group of 28 days to 11 years.

TABLE II - (Distribution of different Cause of death and counts)

S.No	**CLASS/TARGET**	**Agincourt**	**Matlab**	**All Adult**	**India Adult**	**All Child**	**India Child**
1	Acute_Respiratory:	110	11	304	81	532	141
2	Neonatal Conditions	NA	NA	NA	NA	NA	NA
3	Cardiovascular_ Disecases:	381	714	928	242	76	25
4	Chronic_Respiratory:	27	129	84	52	NA	NA
5	Diarrhoeal:	66	29	101	41	256	112
6	HIV/AIDS:	2012	NA	NA	NA	NA	NA
7	Ill_defined:	711	35	NA	NA	194	65
8	Liver_cirrhosis:	89	100	234	59	NA	NA
9	Maternal:	60	23	345	136	NA	NA
10	Neoplasms(cancer):	244	352	497	19	28	15
11	Nutrition_endocrine:	70	90	NA	NA	NA	NA
12	Pulmonary_TB:	690	43	177	21	NA	NA
13	Road_and_transport _injuries:	219	49	124	32	92	64
14	Suicide:	125	34	70	33	NA	NA
15	other_Non_ Commnicable diseases:	221	244	697	125	186	80
16	other_injuries:	366	68	471	218	324	259
17	other_unspecified _infections:	432	79	622	174	376	187

We applied multinomial logistic regression, InterVA-4, InsilicoVA and Tariff methods on above mentioned datasets and obtained results were compared with Medico assigned cause of deaths. Each dataset mentioned above consisting of target feature with multiple cause of deaths. Table II represents the corresponding cause of death its distribution and count in each dataset. Figure 1 and 2 are shows the distribution of cause of death in the Agincourt, Matlab and India_child dataset where horizontal axis represents the count and vertical axis represents the specific cause of death.

Fig. 1 - Distribution of different causes in the datasets (Agincourt)

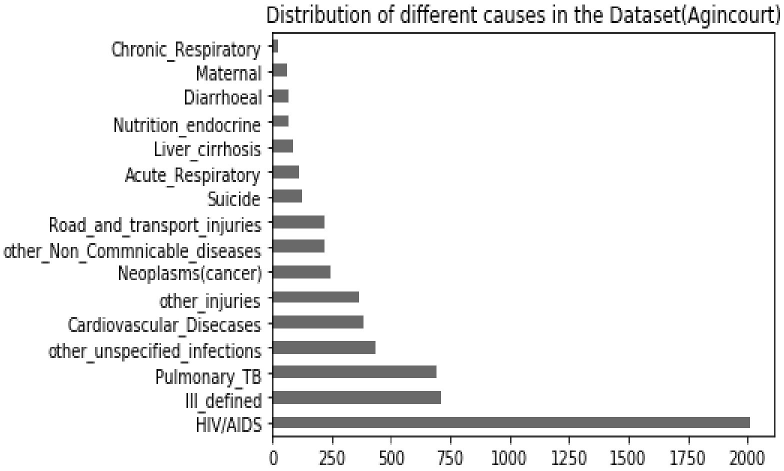
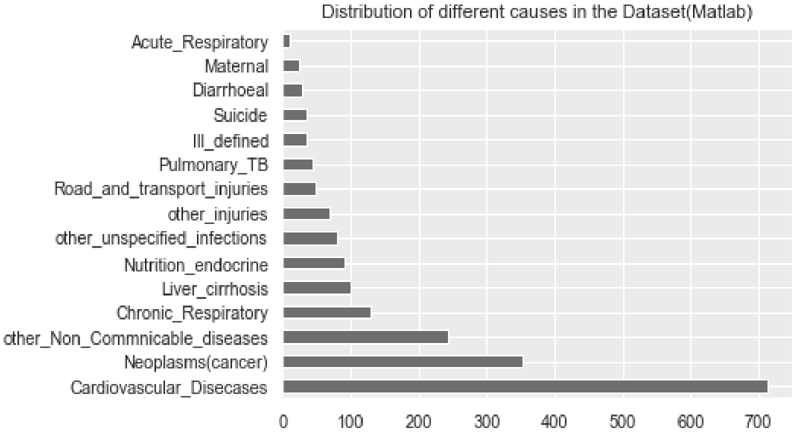


Fig. 2 - Distribution of different causes in the datasets (Matlab)



4. Dimensionality Reduction

Huge dimensionality is curse while we are using some prediction algorithms. Dataset consisting of huge feature can reduces the performance of prediction algorithms. So before applying our prosed model, we used the recursive feature elimination method to reduce number of features from the data. Table III shows the datasets used in this paper along with features before and after dimensionality reduction.

4.1 Recursive Feature Elimination [14]

This is a consequential method for culling the consequential features from the datasets. This method propagates as it is facilely configurable. While configuring this method requires how many numbers of consequential features required and cull of the algorithm. Performance of this algorithm entirely depends on these hyper-parameters. This method can be useful in both classification and regression algorithms. This algorithm utilizes the wrapper type technique to extract the features from the pristine dataset. Wrapper means, this algorithm takes the advantage of another machine learning algorithm for feature selection. In this experiment we used AdaBoostRegressor as wrapper algorithm.

Recursive Feature Extraction Algorithm [16]

```
Tune/train the model on the training set utilizing all P presagers
Calculate model performance
Calculate variable consequentiality or rankings
for each subset size $S_i$ i=1,2,….S do
    Keep the $S_i$ most consequential variables
    [Optional] Pre-process the data
    Tune/train the model on the training set utilizing $S_i$ soothsayers
    Calculate model performance
    [Optional] Recalculate the rankings for each prognosticator
end
Calculate the performance profile over the $S_i$
Determine the felicitous number of soothsayers (i.e. the $S_i$ associated with the best performance)
Fit the final model predicated on the optimal S
```

TABLE III - (Number of features left after RFE)

S.no	Dataset Name	No Of Columns	No of Columns after RFE
1	Agincourt	90	65
2	Matlab	215	181
3	PHMRC_IHME_allSites_Adult_12-69yrs	225	151
4	PHMRC_IHME_India_Adult_12-69yrs	225	151
5	PHMRC_IHME_allSites_Child_28days-11yrs	135	101
6	PHMRC_IHME_India_Child_28days-11yrs	135	101

TABLE III describes the truncation in number of features.

This algorithm initially commences with all the features in the dataset and recursively eliminates the less paramount features while retaining the paramount features.

5. Related Work

The verbal Autopsy dataset consisting of m number of symptoms, n number of deaths; and certain number of causes of death. If the symptom is responsible for cause of death of person, then corresponding entry in the table is marked against $\{s_1, s_2, \ldots . s_m\}$. if the symptom is present in the cause of death of person, then it is marked as 1 or else if the symptom is absent then it is marked as 0 as shown in the Table II. From the Table II $\{c_1, c_2 \ldots c_L\}$ represents the set of cause of deaths.

$$\begin{bmatrix} deaths/sympoms & s_1 & s_2 & & s_{m-1} & s_m & Cause \\ d_1 & 0 & 1 & \cdots & 1 & 0 & c_2 \\ d_2 & 1 & 0 & & 0 & 1 & c_3 \\ \vdots & & & \ddots & & \vdots & \\ d_{n-2} & 0 & 1 & & 1 & 0 & c_3 \\ d_{n-1} & 1 & 1 & \cdots & 0 & 0 & c_4 \\ d_n & 0 & 1 & & 1 & 0 & c_5 \end{bmatrix} \qquad (1)$$

5.1 Tariff [4]

Tariff methods usually depends on the number of symptoms present in the cause of death and its corresponding pattern of score. It computes the tariff score for each cause of death. Tariff score is calculated for cause of death, the cause of death which scores the highest among other is actual cause of death of person.

$$Tariff_{ij} = \frac{v_{ij} - Median(v_{ij})}{Interquartile\ Range\ v_{ij}} \qquad (2)$$

Tariff $_{ij}$ is representing the value computed for ith cause of death for jth symptom, v_{ij} part of Verbal Autopsy represents the response for ith cause for jth symptom. Median (v_{ij}) represents the median part of all positive responses from the ith cause of death for jth symptom. The mean of all cause of death where we get the positive responses was computed by taking the interquartile Range across all positive responses v_{ij}.

$$Tariff\ Score_{kj} = \sum_{j=1}^{m} Tariff_{ij} \vartheta_{jk} \qquad (3)$$

From the equation 2 we are integrating all specific cause of death and corresponding Tariff values computed from the equation 2 with positive responsive value for ϑ_{jk}, where ϑ_{jk} represents the jth symptom for kth death.

5.2 InterVA-4 [5]

Basically, it uses the Bayesian probabilistic modeling to the predict the exact cause of death for a particular human through the Verbal Autopsy data. Bayes theorem for detecting whether a particular event happens or not under unconditional probability. Bayes theorem was used in Verbal autopsy to detect the exact cause of death of particular person. From the Table II we can compute the total conditional probability of jth symptom over the ith cause of death is given by

$$P(c_i|s_j) = \frac{P(s_j|c_i)P(c_i)}{\sum_{i=1}^{L} P(c_i)} \quad (4)$$

5.3 In SilicoVA[6]

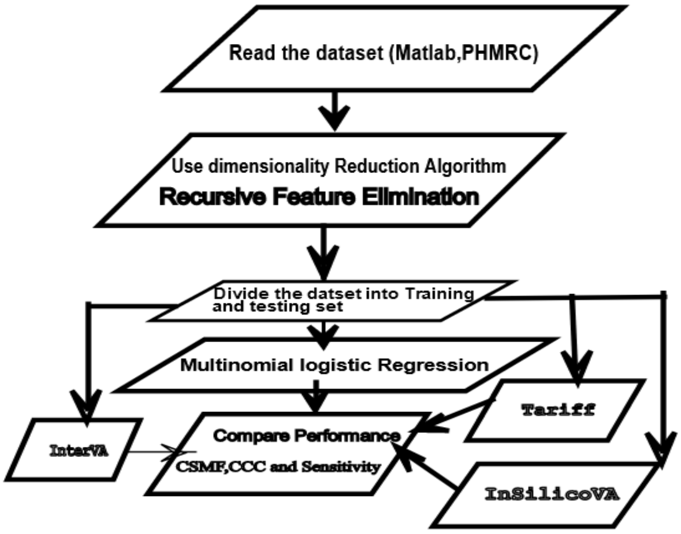
In this model adopted a hierarchical Bayesian framework to estimate the individual cause of death of persons. Their assumption is that each symptom contributed equally and randomly towards the cause of death which follows the Bernoulli distribution. That is

$$s_{ij}|y_i = c \sim Bernoulli(P(s_{ij}|y_i = c)) \quad (5)$$

$P(s_{ij}|y_i = c)$ this expression represents the probability that the symptom s_j is responsible for the given cause of death y_i which is equal to c.

6. Contribution Work

6.1 Methodology


Read the dataset (Matlab,PHMRC)
Use dimensionality Reduction Algorithm
Recursive Feature Ellmination
Divide the datset into Training and testing set
Multinomial logistic Regression
Tariff
InterVA
Compare Performance
CSMF,CCC and Sensitivity
InSilicoVA

Received: 13.10.2021 – Accepted: 08.11.2021

i) Collect and read the original datasets from the WHO (World health organization) and ICD (International Classification of disease).
ii) Use the recursive feature elimination dimensionality reduction algorithm, so data is suitable for processing.
iii) Divide the dataset into training set and test set for; training set is training the model and testing set is for validating and classification.
iv) Multinomial logistic Regression, InterVa-4, InSilicoVA and Tariff methods were trained and testing based on proposed dataset.
v) Compare the performance of these algorithm based on performance metrics like CSMF (Cause specific mortality factor), CCC (Chance corrected Concordance) and Sensitivity.

6.2 Multinomial Logistic Regression [15]

Multinomial logistic regression is kindred to the logistic regression but differences in the dependent target attribute which have more than two classes. Multinomial logistic regression is a relegation technique which is an extension of mundane logistic regression to solve the multi class classification problems. This model usually predicts the probabilities associated with each class in a multi class classification problem.

Following the posits to be made afore applying the multinomial logistic regression.

- The dependent variable is either nominal or ordinal variable.
- The set of independent variables should be either perpetual or ordinal and nominal.
- The set of observations and dependent variables should be independent.

The solutions of multinomial logistic regression is done for K classes we construct K-1 logistic models. Example in our case of dataset Agincourt which has total 16 classes of cause of death for that we construct 15 logistic regression models. Let us consider a problem where dependent variable has 2 classes A and B. Table IV details the cause of deaths in the datasets. Then we take only one logistic regression model with probability of equation as,

Received: 13.10.2021 – Accepted: 08.11.2021

TABLE IV - (List of different Cause of deaths)

Cause
Acute_Respiratory:
Cardiovascular_Disecases:
Chronic_Respiratory:
Diarrhoeal:
HIV/AIDS:
Ill_defined:
Liver_cirrhosis:
Maternal:
Neoplasms(cancer):
Nutrition_endocrine:
Pulmonary_TB:
Road_and_transport_injuries:
Suicide:
other_Non_Commnicable_diseases:
other_injuries:
other_unspecified_infections:

$$Log\left(\frac{p}{1-p}\right) = a + b_1x_1 + b_2x_2 \ldots\ldots + b_nx_n \quad (6)$$

If the value of p>=0.5 then the sample is classified as A or else B.

Lets us take the number of classes in Agincourt dataset from the Table IV. Here n represents the number of attributes in the dataset and x_1, x_2…..x_n are attributes.

$$log\left(\frac{P(\text{Acute\_Respiratory})}{P(\text{other\_unspecified\_infections})}\right) = a_1 + b_1x_1 + \cdots b_nx_n \qquad (7)$$

$$log\left(\frac{P(\text{Cardiovascular\_Disecases})}{P(\text{other\_unspecified\_infections})}\right) = a_2 + b_1x_1 + \cdots b_nx_n \qquad (8)$$

.

.

$$log\left(\frac{P(\text{other\_injuries})}{P(\text{other\_unspecified\_infections})}\right) = a_{15} + b_1x_1 + \cdots b_nx_n \qquad (9)$$

Now the equation 7 and 8 can be rewritten as

$$P(\text{Acute\_Respiratory}) = P(\text{other\_unspecified\_infections}) * \exp((a_1 + b_1x_1 + \cdots b_nx_n)) \qquad (10)$$

$$P(\text{Cardiovascular\_Disecases}) = P(\text{other\_unspecified\_infections}) * \exp((a_2 + b_2x_1 + \cdots b_nx_n)) \ (11)$$

$$P(\text{other\_injuries}) = P(\text{other\_unspecified\_infections}) * \exp((a_{15} + b_2x_1 + \cdots b_nx_n)) \qquad (12)$$

Adding the equations 10,11 and 12 which and sum of probabilities equal to 1.

$$P(\text{Acute\_Respiratory}) + P(\text{Cardiovascular\_Disecases}) \ldots \ldots P(\text{other\_injuries}) = 1 \text{ (13)}$$

Then

$P(\text{other\_unspecified\_infections}) * \exp((a_1 + b_1x_1 + \cdots b_nx_n))+$

$P(\text{other\_unspecified\_infections}) * \exp((a_2 + b_2x_1 + \cdots b_nx_n)) +$

..

$P(\text{other\_unspecified\_infections}) * \exp((a_{15} + b_2x_1 + \cdots b_nx_n))$=1

$$P(\text{other\_unspecified\_infections}) = \frac{1}{1 + \left(\exp\left((a_1 + b_1x_1 + \cdots b_nx_n)\right) + \exp\left((a_2 + b_2x_1 + \cdots b_nx_n)\right) + \cdots . exp((a\_15 + b\_2\ x\_1 + \cdots b\_n\ x\_n\ ))\right)} \quad (14)$$

Once the probability of class Other_Unspecified _insfections is computed the remaining classes can obtained in the same manner.

7. Performance Metrics

Performance of these Verbal Autopsy algorithms and their assessment was done through the metrics like Sensitivity, Chance corrected concordance and Cause specific mortality factor. Medico double optically incapacitated review was taken to from the coding. Here the datasets we divided into 80% training and 20% as testing. We quantified the performance at consummate level and individual level of cause of death.

7.1 Sensitivity [13]

It measures out of all positive values how many of them we have predicted as positive. This metric is useful to measure the how good a model can detect the positives. It also called as Recall.

$$Sensitivity = \frac{TP}{TP+FN} \qquad (15)$$

TP: True Positive and FN: False Negative.

7.2 Chance Corrected Concordance (CCC)[13]

This metric useful to enhance the comparison between different models and their assessment capacity under individual cause of death.

$$CCC_j = \frac{\left(\frac{TP_j}{TP_j+FN_j}\right) - \left(\frac{1}{N}\right)}{1-\left(\frac{1}{N}\right)} \qquad (16)$$

N: Represents the total number of records.

j: Represents the specific cause of death.

This metric gives the negative value for the algorithm assessment when the number of cause of death minimized under 1/N value. The value of CCC is varies between 0 and 1. This metric is additionally serviceable to quantify the performance of different algorithms.

7.3 Cause Specific Mortality Factor (CSMF) [13]

Cause specific moratality factor will assess how closely and accurately the given algorithm will classify the given data and its cause of death.

$$CSMF_{Accuracy} = 1 - \frac{\sum_{j=1}^{L}\left|CSMF_j^{True} - CSMF_j^{Pred}\right|}{2\left(1 - Min\left(CSMF_j^{True}\right)\right)} \qquad (17)$$

From the expression-(17) we can compute the CSMF precision of different cause of death. We can derive the CSMF Precision by taking the distinction between authentic CSMF value and Soothsaid CSMF value of different cause of death of different models. In the equation (17) L represents the number of causes of death classes in the dependent or target attribute.

8. Results and Discussion

Proposed model with its corresponding CSMF and CCC values can be compared to the results of earlier models and studies. Table V shows the values of CSMF and CCC for various models with different datasets. It can be inferred from the table V is CSMF and CCC values of proposed model were higher compared with other models. Table V highlights the percentage of sensitivity at specific cause of death for various models with different datasets. Figure 3 is the graphical summary of various models and their corresponding CSMF and CCC values.

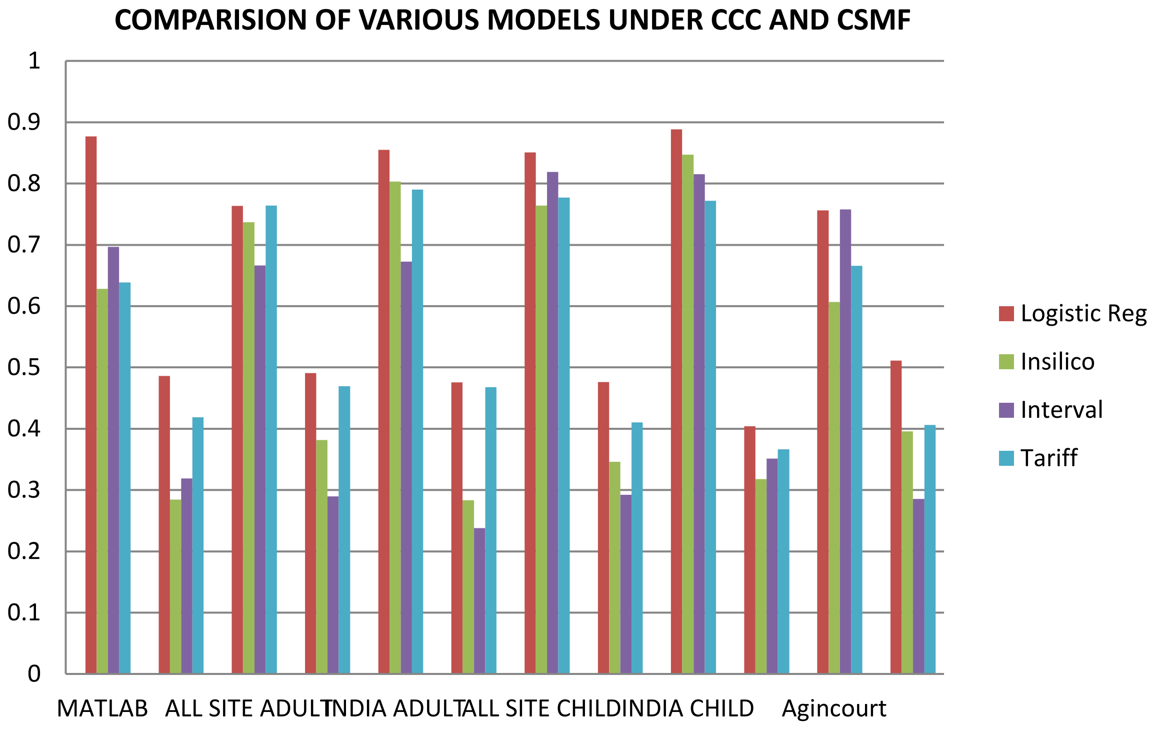
TABLE V - (CSMF and CCC values for various algorithms)

	MATLAB		**ALL SITE ADULT**		**INDIA ADULT**		**ALL SITE CHILD**		**INDIA CHILD**		**Agincourt**	
	CSMF	CCC	CSMF	CCC	CSMF	CCC	CSMF	CCC	CSMF	CCC	CSMF	CCC
Logistic Reg	0.87688	0.48601	0.76362	0.49072	0.85477	0.47565	0.85049	0.4762	0.8883	0.40408	0.75603	0.51107
Insilico	0.62797	0.28441	0.73702	0.38154	0.80302	0.28348	0.76387	0.34586	0.84693	0.31781	0.60665	0.3958
Interval	0.69676	0.31868	0.66611	0.28973	0.67268	0.23793	0.81901	0.2922	0.81508	0.35136	0.75791	0.28561
Tariff	0.63846	0.41862	0.76386	0.46921	0.78992	0.46757	0.77723	0.41019	0.77174	0.36618	0.6658	0.40595

TABLE V1 - (% Sensitivity for the cause of deaths in various datset models)

Cause	**Logistic Regression (Agincourt)**	**Logistic Regression (MATLAB)**	**Logistic Regression (ALL ADULT)**	**Logistic Regression (ALL CHILD)**	**Physician**	**Tariff**	**InterVA-4**	**InsilicoVA**
Acute_Respiratory:	47.7	50	27.2	46.80	1.9	44.3	36.1	53.1
Cardiovascular_ Disecases:	42.7	61.7	53.5	25	6.5	24.7	13.7	14.8
Chronic_Respiratory:	60	68.9	69.2	53.8	0.5	30.8	43.3	35.8
Diarrhoeal:	54.6	40	33.3	NA	1.1	39.8	34.7	31.2
HIV/AIDS:	44.9	NA	NA	NA	34.5	21.4	74.5	29.3
Ill_defined:	19.6	0	NA	33.3	12.2	3	0	29.7
Liver_cirrhosis:	57.2	41.7	59.0	NA	1.5	50	50.7	41.4
Maternal:	100	40	80.9	NA	1	60.3	29.2	52.1
Neoplasms (cancer):	40.9	71.4	75	80	4.2	24.6	28.1	26.2
Nutrition_endocrine:	69.3	42	NA	NA	1.2	69.3	25.8	32.8
Pulmonary_TB:	63.7	75	48.5	NA	11.8	53.3	59.9	60.9
Road_and_transport _injuries:	85.2	100	68.9	88.8	3.8	80.8	78.4	81.5
Suicide:	56.6	83	60	NA	2.1	14	21.9	79.8
other_Non_ Commnicable _diseases:	31.9	23.9	16.5	24.2	3.8	19.6	9.9	18.5
other_injuries:	59.1	72.9	56.0	90.5	6.3	41	64.7	52.7
other_unspecified _infections:	34.1	9.09	40.3	38.2716	7.4	7.2	12.5	11.4

Fig. 3 - Comparison of various models



Over all, this study provides enough evidence and support for the validity of proposed mode. Results obtained by proposed model are consistent with our findings.

9. Conclusion

Automatic Verbal Autopsy is a process of identifying the categorical cause of death by utilizing computer implements and automatic algorithms on Verbal Autopsy data. In generally most of deaths of people transpire at home rather than hospital. Medico predicated autopsy is time consuming and costly process. In this paper we proposed Multinomial logistic regression predicated relegation algorithm to relegate the cause of death. We used Tariff, interval-4 and InSilicoVA anterior methods to compare with our proposed model, categorical metrics like CSMF and CCC are acclimated to compare the performance these algorithms. The evidence from the study and results suggested that proposed model is far better than the anterior models. For further study, we apply incipient methodology and model to enhancing the quality of results.

References

P. Jha, “Reliable direct measurement of causes of death in low-and middle-income countries,” *BMC Med*, vol. 12, no. 1, p. 19, Dec. 2014, doi: 10.1186/1741-7015-12-19.

P. W. Setel *et al.*, “Sample registration of vital events with verbal autopsy: a renewed commitment to measuring and monitoring vital statistics,” *Bulletin of the World Health Organization*, p. 7, 2005.

E. Fottrell and P. Byass, “Verbal Autopsy: Methods in Transition,” *Epidemiologic Reviews*, vol. 32, no. 1, pp. 38–55, Apr. 2010, doi: 10.1093/epirev/mxq003.

S.L. James, A.D. Flaxman, and C.J. Murray, “Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies,” *Popul Health Metrics*, vol. 9, no. 1, p. 31, Dec. 2011, doi: 10.1186/1478-7954-9-31.

P. Byass *et al.*, “Strengthening standardised interpretation of verbal autopsy data: the new InterVA-4 tool,” *Global Health Action*, vol. 5, no. 1, p. 19281, Dec. 2012, doi: 10.3402/gha.v5i0.19281.

T.H. McCormick, Z.R. Li, C. Calvert, A.C. Crampin, K. Kahn, and S.J. Clark, “Probabilistic Cause-of-Death Assignment Using Verbal Autopsies,” *Journal of the American Statistical Association*, vol. 111, no. 515, pp. 1036–1049, Jul. 2016, doi: 10.1080/01621459.2016.1152191.

A.D. Flaxman, A. Vahdatpour, S. Green, S.L. James, and C.J. Murray, “Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards,” *Popul Health Metrics*, vol. 9, no. 1, p. 29, Dec. 2011, doi: 10.1186/1478-7954-9-29.

G. King and Y. Lu, “Verbal Autopsy Methods with Multiple Causes of Death,” *Statist. Sci.*, vol. 23, no. 1, Feb. 2008, doi: 10.1214/07-STS247.

P. Miasnikof *et al.*, “Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths,” *BMC Med*, vol. 13, no. 1, p. 286, Dec. 2015, doi: 10.1186/s12916-015-0521-2.

C. J. Murray *et al.*, “Using verbal autopsy to measure causes of death: the comparative performance of existing methods,” *BMC Med*, vol. 12, no. 1, p. 5, Dec. 2014, doi: 10.1186/1741-7015-12-5.

P. Byass, Dao Lan Huong, and Hoang Van Minh, “A probabilistic approach to interpreting verbal autopsies: methodology and preliminary validation in Vietnam,” *Scand J Public Health*, vol. 31, no. 62_suppl, pp. 32–37, Dec. 2003, doi: 10.1080/14034950310015086.

P. Serina *et al.*, “Improving performance of the Tariff Method for assigning causes of death to verbal autopsies,” *BMC Med*, vol. 13, no. 1, p. 291, Dec. 2015, doi: 10.1186/s12916-015-0527-9.

S.S. Murtaza, P. Kolpak, A. Bener, and P. Jha, “Automated verbal autopsy classification: using one-against-all ensemble method and Naïve Bayes classifier,” *Gates Open Res*, vol. 2, p. 63, Jan. 2019, doi: 10.12688/gatesopenres.12891.2.

I. Guyon, J. Weston, S. Barnhill, and V. Vapnik, “gene selection for cancer classification using support vector machines,” *Machine Learning*, vol. 46, no. 1/3, pp. 389–422, 2002, doi: 10.1023/A:1012487302797.

J. M. Hilbe, *Logistic Regression Models.* Hoboken: CRC Press, 2009.

M. Kuhn and K. Johnson, *Applied predictive modeling*. New York: Springer, 2013.

Author's Profile

Zainab Mohanad Issa, she is a research scholar in Computer science and Engineering Department, in Acharya Nagarjuna University, Guntur, A.P, India. She has M.Sc. (Information system) in 2017 from Osmania university, Hyderabad, India. She is research active in the area of bigdata& data mining and machine learning.

Dr. Shaheda Akthar, received Bachelor of Computer Science and Master of Computer Science from Acharya Nagarjuna University, M.S from B.I.T.S Pilani. Ph.D from Acharya Nagarjuna University. Presently working as Head of the department of Computer Science, Government College for Women (A), Guntur, A.P, India and Research Director for Department of Computer Science & Engineering, Acharya Nagarjuna University, Guntur, A.P, India. Areas of interest are Software Engineering, reliability and quality control, Software Architecture Recovery. Machine Learning and Data Mining. Published more than 35 research papers in various international journals.

Received: 13.10.2021 – Accepted: 08.11.2021