North Asian International Research Journal of Sciences, Engineering & I.T.

IRJIF. I.F. : 3.821

ISSN: 2454-7514 Vol. 9, Issue-12 December-2023

Index Copernicus Value: 52.88 Indian Citation Index Thomson Reuters ID: S-8304-2016

A Peer Reviewed Refereed International Journal

DOI: 10.5949/nairjseit.2023.10.12.3

ENHANCED SELECTED FEATURE BASED CATEGORICAL BOOSTING CLASSIFIER: IMBALANCED AUTOMATIC VERBAL AUTOPSY CLASSIFICATION

[1]DR.SHAHEDA AKTHAR

[1]Post-Doctoral Fellowship Scholar, Department of Computer Science and Engineering,
Central Christian University,
MALAWI.
[1]Lecturer in Computer Science, Government College for Women (A), Guntur &Research Supervisor, Dept. of Computer Science and Engineering, Acharya Nagarjuna University, Guntur.
Email: shahedaakthar76@gmail.com

ABSTRACT

Before medical professional bodies declare the cause of death, a verbal autopsy is one of the greatest medical techniques for detecting the cause of death automatically. The process of establishing the exact explanation is time-consuming and unclear. To develop forecasts about people's lives and medical services, every country requires a record with an actual cause of death. Standard datasets like PHMRC and Matlab were used, and they were well-received. These datasets, however, featured imbalanced records, making categorization problematic. As a consequence, we employed the Synthetic Minority Oversampling Technique to fix the unbalanced features, as well as dimensionality reduction to get rid of the extra ones. Finally, we classified the data using a Cat Boost Ensemble method. We used three popular metrics to compare our model to previous models like Insilico VA, Tariff, and InterVA-4: sensitivity, chance corrected Concordance (CCC), and cause-specific mortality factor (CSMF). The findings suggest that the proposed model outperforms several other models.

***KEYWORDS**: cause of death classification, Multi class linear discriminant analysis, Chance Corrected Concordance (CCC), Cause-specific mortality fraction (CSMF).*

INTRODUCTION

Determining the major cause of death is frequently a lengthy and complicated process. On the inside, there is no conventional technical approach for diagnosing the particular cause of death. In poor nations, verbal autopsy investigations are increasingly being used for disease surveillance and sample registration to estimate cause-specific mortality. Death is more likely in undeveloped nations when people die outside rather than in hospitals. Using WHO

(World Health Organization) procedures and ICD (International Classification of Diseases) codes, two competent medicos do a verbal autopsy (international relegation of Diseases). These two Medicos amass information on a person's death by questioning his close relatives. If there are any conflicts, the topic can be brought to the attention of the senior Physician to be resolved. There are disagreements over the results of the manual autopsy.

In order for governments to adopt health-related legislation and regulations, verbal autopsies are quickly becoming a critical component in ascertaining the real cause of death. The use of physician-based cause-of-death classification has been challenged in the past, and it is also costly. National and international legislators, public health authorities, and medical professionals require statistics on the worldwide distribution of deaths by cause in order to define research goals, budgetary priorities, and efficacious measures.

The most common method has been a physician's evaluation of symptoms without a validation sample up until now. This procedure is expensive since each death requires three physicians, each of whom takes 20–30 minutes to analyze symptoms and classifications. To reduce the overall time necessary, more physicians might be hired and collaborate.

Physicians must be from the area since their decisions are highly influenced by their past experiences (a Kansas doctor hearing "fever and vomiting" would not think of malaria). This can lead to difficult logistical concerns, as physicians are typically in short supply in these areas, as well as major ethical issues, as doctors are forced to treat patients on the spot. These algorithms operate on datasets that have previously existed in the form of verbal autopsies. In the last several years, Automatic Verbal Autopsy has gotten a lot of press. Automatic Verbal Autopsy is generating a lot of buzz because of its versatility in terms of technique application and time saving. The approach has also been used to investigate sickness risk factors, outbreaks of infectious diseases, and the effect of public health measures.

Verbal autopsy is anticipated to become a crucial component in the medical field. There is a large amount of literature on Verbal Autopsy. [1] Established a link between reliable cause-of-death data and disease-control priorities, enabling for the early discovery of possible epidemics. [2] proposed a cost-effective and nationally representative design for sample vital registration. [3] In his study, he performed an excellent job interpreting verbal autopsy data and diagnosing the cause of death, and he was particularly focused on keeping structured, quantitative, and qualitative records in a biologically sound framework.

[4] Constructed an experimental method known as Tariff for Automatic Verbal Autopsy and found that it is transparent, flexible, and easy to use. It has also passed extensive testing. For the purpose of interpreting Verbal Autopsy data that conforms to the International Classification of Diseases version, a new probabilistic model for InterVA-4 [5] has been developed. [6] A statistical technique known as InsilicoVA was designed to balance the distribution of population causes of death with the distribution of individual cause of death distributions using a data augmentation strategy. (7) Predicting and differentiating the cause of death was done using the machine learning-based Random Forest approach.

[8] A more computational approach was presented here, eschewing Medicos's postulations, expert algorithms, and parametric statistical postulations in favor of concentrating just on the data's theoretical implications and empirical analysis. In [9] We applied the Nave Bayes Classifier in Automatic Verbal Autopsy and compared the outcome with other Medico-predicted demotion. Several models were assessed for validity in [10], and it was discovered that Tariff, Simplified symptom pattern, Random-Forest, Traiff Method, King-Lu, and Medico evaluation of VA forms work better than InterVA-4. [11] In this study, the author has investigated and studied the operation of a probabilistic method of Bayesian probability model for Automatic Verbal Autopsy. [12] In order to fix the shortcomings of Tariff

1.0 and create Tariff 2.0, the author of this paper examined the inner workings of Tariff 1.0 and applied rigorous methodologies [18,19.20].

Reducing the dimensionality of the data while using a classification algorithm is the aim of this study. We addressed the issue of class imbalance by utilizing the CatBoost ensemble method, recursive feature elimination (RFE) for dimensionality reduction by filtering the felicitous features, and the Synthetic Minority Oversampling Technique, which is an excellent method for classification. In the second section, we go over the datasets used for the computation. Section III covers the topics of deep learning, recursive feature elimination, and the synthetic minority oversampling technique. In Section IV, the metrics used to evaluate the performance of different models are covered in more detail. The results of the various models are examined in Section V. See Section VI for the conclusion.

DATA SETS USED

To assess how well the suggested model performed in comparison to earlier[9][13] techniques, we used Verbal Autopsy datasets from two demographic surveillance locations in Agincourt, South Africa [Kahn K, Collinson] and Matlab, Bangladesh [Matlab]. The datasets and the corresponding properties are summarized in Table I. The remaining data was submitted by the 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	PHMRC_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

The Agincourt dataset, which comprises 90 columns and 5823 rows with double clinical coding, is shown in TABLE I. Knowledge The doctor supplied a Matlab dataset that was obtained using single coding and included 2000 rows and 215 columns. PHMRC predicted dataset contains features for children and adults ranging in age from 28 days to 11 years.

We applied Deep learning, InterVA-4, InsilicoVA, and Tariff methods on the previously described datasets, and compared the outcomes with Medico's stated cause of death. There is a goal feature with different reasons of death in each of the datasets described above. Table II provides the distribution and count of the associated cause of death for each dataset. Figures 1 and 2 depict the cause of death distribution in the Agincourt, Matlab, and India child datasets, with the vertical axis denoting the exact cause of death and the horizontal axis showing the count.

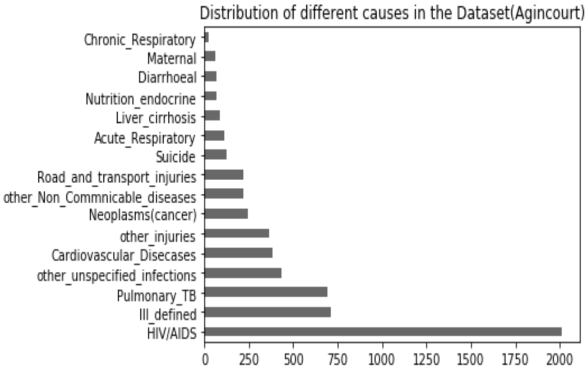

Fig. 1 Distribution of different causes in the datasets (Agincourt).

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

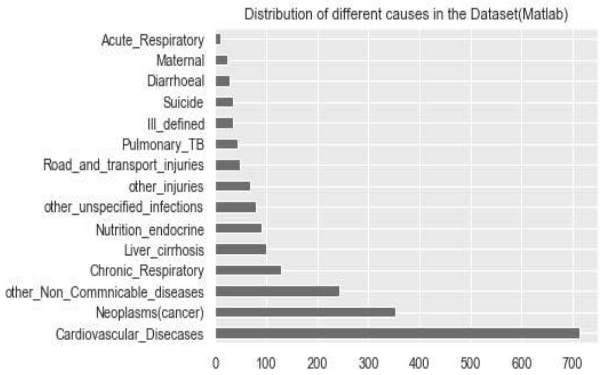

Fig. 2 Distribution of different causes in the datasets (Matlab).

TECHNICAL FRAME WORK

Recursive Feature Elimination [14]

This method works well for eliminating significant features from datasets. This method becomes popular because it is easy to set up. When setting up this process, you will need to know how to filter the algorithm and how many consequential qualities you will require. These hyper-parameters determine the algorithm's performance completely. This approach is useful for both classification and regression techniques. This method makes use of a wrapper-type strategy to extract features from the clean dataset. As a wrapper, this approach suggests that it selects features using a different machine learning technique. RFE [14] selects predictors in reverse.The first step in this strategy is to build a model using all of the predictors and give each one a significance score. After the least significant predictor or predictors are excluded, the model is rebuilt and significance scores are determined once more.

The quantity and composition of predictor subgroups to be assessed are, in actuality, decided by the analyst. Thus, the subset's size serves as an RFE tuning parameter. Predictor selection is based on the subset size that maximizes the performance criterion and the predictors' relevance rankings. Next, the optimal subset is used to train the final model. In this experiment, we used the wrapper algorithm AdaBoost Regressor.

Recursive Feature Elimination Algorithm[14, 15]

```
1 Tune/train the model on the training set using all $P$ predictors
2 Calculate model performance
3 Calculate variable importance or rankings
4 for each subset size $S_i$, $i = 1. . .S$ do
5     Keep the $Si$ most important variables
6     [Optional] Pre-process the data
7     Tune/train the model on the training set using $Si$ predictors
8     Calculate model performance
9     [Optional] Recalculate the rankings for each predictor
10 end
11 Calculate the performance profile over the $S_i$
12 Determine the appropriate number of predictors (i.e. the $S_i$associated with the best performance)
13 Fit the final model based on the optimal $S_i$
```

TABLE III (Number of features left after RFE)

Sno	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 shows the number of characteristics that have been truncated. This approach starts with all of the characteristics in the dataset and iteratively removes the less important features while keeping the important ones.

Handling Imbalancing Features [16]

Machine learning algorithms' performance is often assessed using predictive accuracy. But when the data is imbalanced and/or the costs of certain errors are highly variable, this is inefficient. The field of machine learning has adopted two strategies to address the problem of class imbalance. Assigning a different price to each training example is one way to go. Re-sampling the original datasets, either by over- or under-sampling the minority class, is the second option. In this article, we used an over-sampling method whereby the minority class was over-sampled without replacement by creating "synthetic" samples.This methodology is based on an effective handwritten character recognition system [17]. They produced more training data by executing specific operations on actual data. In their case, rotation and skew were reasonable methods to slant the training data. By working in "feature space" instead of "data space," we can create less application-specific synthetic instances. Over-sampling of the minority class is achieved by introducing synthetic cases along the line segments linking any/all of the k nearest neighbors.

Cat Boost (Categorical Boosting)

An approach to building a weighted ensemble is called boosting. The base algorithms are added one by one. N classifiers are learned iteratively. Adjusting weights enables subsequent classifiers to "pay greater attention" to training tuples that were incorrectly classified in the past.Dorogush created the machine learning method known as "categorical boosting," or "CatBoost," which uses gradient boosting on decision trees [18].

An open-source gradient boosting tree library called CatBoost was developed by Yandex engineers and academics. As its name suggests, this package is capable of handling categorical attributes. The simplest method for handling categorical characteristics is to convert them to one-hot encoding.

With the ability to specify categorical column indices, CatBoost facilitates one-hot encoding with one-hot-max-size (Use one-hot encoding for any features with the number of distinct values less than or equal to the given parameter value). If the category features parameter has no features, CatBoost will read all columns as numerical variables. CatBoost uses an effective encoding method similar to mean encoding, but it minimizes over-fitting for the remaining categorical columns whose unique number of categories is greater than one-hot-max-size. This is how the process operates.

1. Creating a random order for the collection of incoming observations.

A large number of random permutations are created.

2. Converting a floating-point or category label value to an integer.

3. The values of all category features are converted to numeric values.

PERFORMANCE METRICS

Sensitivity, Cause specific mortality factor, and Chance adjusted concordance were employed to evaluate and estimate the efficacy of these Verbal Autopsy techniques. The code was entered for a physician's evaluation of double visual impairment. The datasets were divided into training and testing categories. Both globally and personally, we evaluated the cause of death's performance.

Sensitivity [13]

It determines the proportion of all positive values that we expected to be positive. A model's ability to identify false positives may be evaluated using this statistic. Recall is another name for it. The sensitivity number indicates how well our model predicts true positives when it is high. The true positives-to-all-positives ratio can be defined as the ratio of true positives to all positives.

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP}+\text{FN}} \qquad (1)$$

TP: True Positive and FN: False Negative.

Chance Corrected Concordance(Ccc)[13]

This statistic can be used to compare various models and their capacity to identify particular causes of death in a range of situations.

$$\text{CCC}_j = \frac{\left(\frac{\text{TP}_j}{\text{TP}_j+\text{FN}_j}\right)-\left(\frac{1}{N}\right)}{1-\left(\frac{1}{N}\right)} \qquad (2)$$

N :represents the total number of records.

j: represents the specific cause of death.

This metric yields a negative algorithm evaluation rating when the number of causes of death is minimized below the 1/N ratio. The value of CCC is between 0 and 1. It is also possible to compare the performance of several algorithms using this statistic.

Cause Specific Mortality Factor(Csmf)[13]

The degree to which the suggested algorithm accurately and closely classifies the data and the cause of death will depend on the cause-specific mortality factor.

$$\text{CSMF}_{\text{Accuracy}} = 1 - \frac{\sum_{j=1}^{m}\left|\text{CSMF}_j^{\text{True}} - \text{CSMF}_j^{\text{Pred}}\right|}{2\left(1-\text{Min}\left(\text{CSMF}_j^{\text{True}}\right)\right)} \qquad (3)$$

The CSMF precision of various causes of death can be calculated using expression (3). The difference between the actual and Soothsaid CSMF values of various causes of death for various models can be used to calculate the CSMF Precision.

Datasets	**ALL SITE**	**INDIA**	**ALL SITE**	**INDIA**
models	**ADULT**	**ADULT**	**CHILD**	**CHILD**
CatBoost	0.7699	0.8131	0.7525	0.8236
InSilicoVA	0.4195	0.6809	0.4769	0.5098
Tariff	0.5509	0.6554	0.5073	0.5533

Table IV Sensitivity comparisons of different models

RESULTS AND DISCUSSION

The CSMF and CCC values of the suggested model can be compared to those of other models and studies. The CSMF and CCC value for a range of models and datasets are shown in Table V. Table V indicates that the CSMF and CCC values of the proposed model were higher than those of the other models. Table IV displays the percentage of sensitivity for multiple models using various datasets related to certain causes of death. A graphical illustration of many models together with the corresponding CSMF and CCC values can be found in Figure 4.

	MATLAB		**ALL SITE ADULT**		**INDIA ADULT**		**ALL SITE CHILD**		**INDIA CHILD**		**Agincourt**	
Model	CSMF	CCC	CSMF	CCC	CSMF	CCC	CSMF	CCC	CSMF	CCC	CSMF	CCC
CatBoost Ensemble	0.9548	0.8774	0.9397	0.7507	0.9575	0.7975	0.9701	0.7215	0.941	0.8015	0.9375	0.7933
Insilico	0.7992	0.4673	0.7852	0.3711	0.875	0.5763	0.8648	0.4115	0.8231	0.4485	0.8433	0.4165
Interval	0.7167	0.4592	0.6256	0.3067	0.7314	0.5034	0.7694	0.3997	0.8373	0.5416	0.6196	0.3058
Tariff	0.837	0.5503	0.8995	0.5135	0.868	0.6267	0.915	0.4457	0.8725	0.4975	0.9089	0.4462

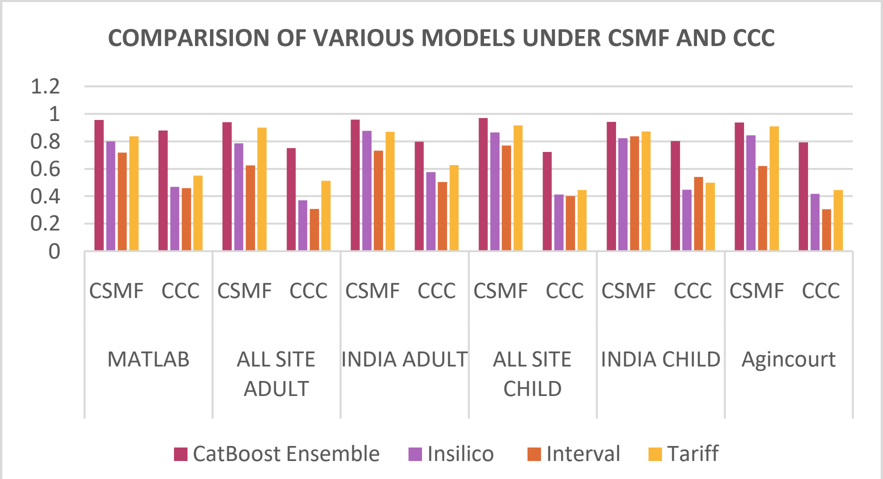

Fig.3 Comparision of various models

Overall, there is sufficient data in this investigation to validate the suggested mode. Our findings are in line with the outcomes of the suggested model.

CONCLUSION

Automatic Verbal Autopsy is the process of identifying the categorical cause of death from Verbal Autopsy data utilizing computer technology and automatic algorithms. Rather than in hospitals, most fatalities take place at home. A medico-predicted autopsy is a costly and time-consuming process. In this study, we proposed an ensemble-based Cat-Boosting classification of mortality causes. Using real-time datasets from Agincourt, Matlab, and PHMRC, SMOTE (Synthetic Minority Oversampling Technique) and recursive feature elimination were used to reduce dimensionality and address the issue of class imbalance. We used the Tariff, interval-4, and Insilico VA prior approaches to examine the performance of these algorithms. The effectiveness of these algorithms was evaluated using classification measures such as CSMF and CCC. The suggested model outperforms the earlier models by a significant margin, according to the study's data and findings. We employ an effective methodology and model to raise the caliber of the output.

REFERENCES

[1] 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.

[2] P. W. Setelet 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.

[3] 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.

[4] 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.

[5] P. Byasset 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.

[6] 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.

[7] 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.

[8] 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.

[9] P. Miasnikofet 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.

[10] 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.

[11] P. Byass, Dao LanHuong, 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.

[12] 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.

[13] 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.

[14] 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.

[15] M. Kuhn and K. Johnson, Applied predictive modeling. New York: Springer, 2013.

[16] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic minority over-sampling technique," Journal of Artificial Intelligence Research, vol. 16, pp. 321–357, Jun. 2002. Accessed: Feb. 10,

2022.[Online]. Available: https://doi.org/10.1613/jair.953

[17] T. M. Ha and H. Bunke, "Off-line, handwritten numeral recognition by perturbation method," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 19, no. 5, pp. 535–539, May. 1997. Accessed: Feb. 10, 2022. [Online]. Available: https://doi.org/10.1109/34.589216

[18] V. Dorogush, V. Ershov, A. Gulin, CatBoost: gradient boosting with categorical features support, arXiv preprint arXiv:1810.11363 (2018)

[19] ZMI Ansaf, ShahedaAkthar: revistageintec-gestaoinovacao e tecnologias 11 (4), 5857-5872Automatic Verbal Autopsy Classification Using Multinomial Logistic Regression Classifier by Using Recursive Feature Elimination

[20] Dr. Shaik Mohammad Rafi, Z. M. I. A. D. ShahedaAkthar A. (2021). Automatic Verbal Autopsy Classification Using Multi-Class Linear Discriminant Analysis and Recursive Feature Elimination. Design Engineering, 3949-3964. Retrieved from http://www.thedesignengineering.com/index.php/DE/article/view/5344

[21] ZMI Ansaf, ShahedaAkthar:Advances in Mechanics Volume 9, Issue 3, 2021 Page 1382- 1388 Automatic verbal autopsy classification using hard voting classifier: recursive feature eliminatio