

Robust Application of Supervised Machine Learning Techniques for Cause of Death Determination from Verbal Autopsies

Michael T. Mapundu (✉ michael.mapundu@wits.ac.za)

School of Public Health, Department of Epidemiology and Biostatistics, University of The Witwatersrand, Johannesburg, South Africa

Chodziwadziwa W. Kabudula

MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt)

Eustasius Musenge

School of Public Health, Department of Epidemiology and Biostatistics, University of The Witwatersrand, Johannesburg, South Africa

Victor Olago

National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg, South Africa.

Turgay Celik

School of Electrical and Information Engineering, University of The Witwatersrand, Johannesburg, South Africa.

Research Article

Keywords: machine learning, natural language processing, verbal autopsy

Posted Date: November 8th, 2021

DOI: <https://doi.org/10.21203/rs.3.rs-1010158/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

RESEARCH

Robust Application of Supervised Machine Learning Techniques for Cause of Death Determination from Verbal Autopsies

Michael T. Mapundu^{1*}, Chodziwadziwa W. Kabudula^{1,2}, Eustasius Musenge¹, Victor Olago³ and Turgay Celik^{4,5}

*Correspondence:

michael.mapundu@wits.ac.za

¹ School of Public Health,,
Department of Epidemiology and
Biostatistics, University of The
Witwatersrand, Johannesburg,
South Africa

Full list of author information is
available at the end of the article

Abstract

Background: Verbal Autopsy (VA) is a tool commonly used in low to medium income countries to ascertain cause of death, where most deaths are not assigned a medically certified cause. As such, they strengthen health priorities, inform policy, practice and provide vital information where civil registration systems are weak. Physician diagnosis is used as a gold standard to determine cause of death, from VA interviews even though it is inconsistent and expensive. Alternatively, conventional computer algorithms and machine learning approaches have been applied. However, they fail to perform optimally because of data quality and ineffective strategies that they employ. We present a robust machine learning framework that can accurately classify cause of death using only narratives from VA interviews.

Methods: Experiments started with data acquisition of the VA narratives, followed by data preprocessing. We created numeric vectors to represent the narratives using various feature engineering techniques for twelve cause of death categories. Furthermore, we applied data balancing, feature scaling, hyper-parameter tuning and dimensionality reduction in order to improve model performance. We applied eight different classification approaches to the vectors to generate model predictors of cause of death. Validation was done using Precision, Recall, Accuracy, F1-score and Receiver Operating Characteristic Area Under Curve (ROCAUC).

Results: We used the physician diagnosis as our gold standard for validation of our models. Our five best classifiers attained a Precision, Recall, Accuracy and F1-score of 95%, 94%, 93%, 92% and 91% respectively in cause of death classification of all twelve disease categories. We report on Micro-Average ROCAUC of 96% and Macro-Average ROCAUC of 95% of our twelve classes.

Conclusion: Our proposed robust machine learning framework can be a faster and cost effective way to determine cause of death from rich informative unstructured VA narratives. This study can also serve as a benchmark of model comparability and generalisation of machine learning models in determining cause of death using VA data. Our study was limited in terms of data quality. Future work aims at using combined responses and narratives for our models and also applying deep learning architectures for cause of death classification using VAs.

Keywords:

machine learning; natural language processing; verbal autopsy

Background

Most of the countries in the world fail to meet the United Nations 90 percent death registration coverage requirement, as deaths in many Low to Medium Income Countries (LMICs) are not captured in civil registration systems [1, 2]. In addition, 65 percent of the population in the world lack high quality information on cause of death since every year about sixty million deaths worldwide are not assigned a medically certified cause [3]. However, the cause of death information is important as it is needed for public health monitoring, to inform critical health policies and priorities. As such, in the absence of clinically oriented sources, cause of death information should be derived from alternative sources. The most common approach used globally as an alternative source of cause of death information is Verbal autopsy (VA). VA is a process that is used to determine the cause of death where death occurs outside health facilities and is not certified by a doctor and is common in LMICs [4]. The VA process is done by a non-clinical personnel who conduct interviews using a structured questionnaire with relatives of the deceased about circumstances and events that led to death, and this information is captured as text [3]. The VA process suffers from inconsistencies and inaccuracies as it is subjective, prone to errors amongst many drawbacks. The compiled VA narratives are then given to two doctors for assessment who reach a consensus on the cause of death and if not a third physician is consulted a process known as Physician Coded Verbal Autopsy (PCVA). PCVA is the only available gold standard for determining cause of death even though it is widely criticised because of lack of robustness, cost and time amongst many drawbacks [5]. Therefore, PCVAs are mostly employed for training and validation of computational approaches. Advances in technology have given rise to automated methods for determining cause of death which are faster, efficient and cost effective [3]. In this study we present a robust machine learning framework for determining causes of death only from VA narratives. We apply effective data cleaning strategies, data balancing to achieve optimum transparency and accuracy through addressing most model limitations and applying recommendations that are reported in [6, 7, 8]. We assess the robustness of several classifiers including; random forest (RF), k-nearest neighbour (KNN), decision tree (DT), support vector machine (SVM), logistic regression (LR), artificial neural network (ANN), Naive Bayes (NB) and bagging as an ensemble classifier.

Computer Coded Verbal Autopsy

Literature to date reports on five various Computer Coded Verbal Autopsy (CCVA) or VA algorithm approaches which are expert-driven to perform cause of death classification [9, 10, 11, 12, 13]. These approaches make use World Health Organisation standardised VA instrument which is a questionnaire with questions on signs and symptoms of respondent's health history. The CCVA algorithms are made up of VA data derived from real deaths, symptom-cause information (SCI) which is a repository of information about symptoms that are related to each probable cause of death. This information is derived from the physician's reports. Additionally, they make use of logic that entails a logical algorithm that combines the tSCI and VA data to identify cause-specific mortality fractions (CSMF), so as to assign specific cause of death.

The InterVA algorithm uses the Bayes Rule, which is a function that computes conditional probabilities of a sample when given evidence and Symptom Cause Information (SCI) as conditional probabilities of symptoms given a specific cause of death using CSMFs [14]. [13] applied the InterVA-4 and reported a Sensitivity of 43% and CSMF Accuracy of 71% using data from the Million Death study.

The Tariff algorithm uses fewer symptoms and it employs the SCI as a tariff score to rank causes of death, thus determining the association between specific symptoms and causes in the Population Health Metrics Research Consortium (PHMRC) gold standard dataset. CSMFs are arrived at by identifying a single cause with the highest rank for each death in the VA dataset and sum them up [15]. [10] applied the Tariff approach on the PHMRC and reported 50.5% Chance-corrected concordance (CCC) (a measure of how well the predicted cause of death categories correspond to the correct cause of death categories) and CSFM accuracy of 77%.

The InSilicoVA algorithm uses a statistical approach employing joint probabilities to identify the most likely significant cause of death in relation to CSMFs for all deaths in a VA data set. The study of [12] applied the InSilico approach and reported a mean Sensitivity of 34.1% across 34 cause of death categories and CSMF accuracy of 85% CSMF.

Naïve Bayes Classifier algorithm uses the Bayes Rule to categorise cause of death. The study of [13] applied the NBC and achieved better results as compared to InterVA and the Tariff approach with a Sensitivity of 57% and CSMF Accuracy of 88%. Nevertheless, they used various datasets for model evaluation targeting 16 cause of death categories and their model only used data from the structured questionnaire.

The King-Lu algorithm uses symptom conditional probabilities to estimate cause of death of a dataset over 13 categories. It does not provide a cause of death for individual records [3]. This algorithm relies on SCI training data which defines clusters of symptoms rather than a single symptom. Moreover it is recommended to use gold standard deaths and if possible they should be from the same population as the VA deaths to get credible and better results [15]. [9] used the King-Lu on the Indian Million Death Study Dataset and reported a CSMF Accuracy of 96%.

[16] did an investigation that focussed on validation of VA expert algorithms and concluded that population level accuracy is similar to that of machine learning approaches with CSFM in the range of 57% – 96%. These findings are supported by [17] who did a study where they validated data derived algorithms against the gold standard of physician review and concluded similar findings for certain disease categories based on the CSFM.

Another validation study done by [18] where they re-assigned the VA review process to other physicians to do the diagnosis of cause of death given the VA narratives using the Tariff, Simplified Symptom Pattern (SSP), InterVA and the Random Forest. They got a CSFM in the range of 76.4% – 77% for adults as compared to PCVA that attained 68% and InterVA 62.5% respectively. On the other hand, they report a CSFM Tariff score of 78.3% for children as compared to PCVA that attained 67.8% and InterVA 52% respectively. Similar comparisons are reported in the work of [19], where they compared PCVA and CCVA in determining cause of death and they reported an overall chance concordance of less than 50%.

They concluded that there is little evidence to justify the CCVA as a possible replacement of the gold standard which is the PCVA. Therefore, there is need for further investigations and research with large datasets to train and test models on cause of death classification.

However, there has been less research done in using Machine Learning (ML) which make use automated computer programs that can take data and learn new trends and patterns for VA classification. They also employ a performance measure or weighting to improve the performance. Moreover, it employs some statistical, probabilistic and optimisation technique in order to discover patterns and trends in complex data through some analysis to get to a decision [20].

Machine Learning in VA

[3] argues that to date, ML techniques have been primarily applied to data from the structured questionnaires only, with the best Sensitivity scores around 60% for individual cause of death classification, using various numbers of cause of death categories. ML can avail real-time results that is similar to that of physicians/experts [21]. In literature, complex machine learning (ML) models can be found that can replace the PCVA and CCVA algorithms as approaches of determining cause of death.

[22] used the LR model to determine the completion rate of VA and factors associated with undetermined cause of death. They reported a 83% to 89% completion rate. [23] reports various common diseases that lead to death using CSFM and LR classifier and they achieved 80% Specificity. [24] applied ANN classify cause of death from VAs and achieved a Sensitivity of 45.3. On the contrary, this is the only study that has used ANN to date and they concluded that more explorations are needed with large datasets and large training samples to improve results of the ANN. We also explore with the ANN in this study.

[25] used the RF classifier to assign cause of death categories and reported that the algorithm performed better if not as the PCVA approach. They further state that the RF was better than PCVA on overall chance concordance and CSFM accuracy for both adults and children. [26] used text classification techniques to predict cause of death from forensic autopsy reports and found out that the SVM produced better results as compared to other classifiers with Precision of 78.1%, Recall of 78.3%, F-score of 78.2% and overall Accuracy of 78.25% for 16 disease categories. In a similar study done by [27], SVM also outperformed the other classifiers. This might be attributed to the fact that SVM can easily handle non-linearity of data and its capability to handle overfitting. Similar findings are reported in [28] where SVM achieved an accuracy of 95.41%. [29] did a study where they performed automatic classification of diseases using VA narratives using SVM and they reported an F-score in the range of 80%–96%. They deduce that feature extraction approaches are grossly affected by variations in words as well as word combinations. As such, SVM performs better in high dimensional feature spaces because of feature independence and its ability to handle non-linearity.

[30] did a semantic analysis on infants VA data on cause of death in trying to showcase the relationship between keywords and a given cause of death. Furthermore, they did another investigation on linguistic features using infant data from

Ghana and sought to classify into 16 target classes. They achieved a Sensitivity of 40.6% using features from narratives and 61.6% using both structured questionnaires and VA narratives. They conclude that using word occurrences produced better results as compared to word occurrence features. More explorations with large datasets in the medical domain with effective model training might improve model performance. The same authors developed a VA corpus using a structured VA questionnaire and argue that there are properties of the human languages in a VA corpus that are the same with other databases [31].

[32] did some work using Term Frequency with Inverse Document Frequency (TF-IDF) automated classification to determine cause of death from VAs and they proved that TF-IDF is an optimal vectoriser that can improve model performance and if integrated in model design and production, it can improve cause of death categorisation.

[33] did a multi class classification study to determine accident related cause of death using expert driven feature selection. They achieved an evaluation measure of 85%–90% on the RF and DT classifier. Furthermore, their models improved on Accuracy by 14% to 16%. The models outperformed the SVM, NB and KNN. [28] did a study where they classified VA reports based on conceptual graph-based method document representation model and they report a 12%-15% model improvement on performance as compared to fully automated baseline graph based document representation techniques. [26] used text classification techniques to classify cause of death from forensic autopsy reports. They report that uni-grams are better feature extraction techniques, Term Frequency (TF) and TF-IDF being better feature representation schemes. They further state that Chi-squared produced better performance as a dimensionality reduction approach. On classifiers they point out that the SVM outperformed RF, NB, KNN, DT and ensemble classifiers. [8] did a systematic survey on current literature of clinical text mining techniques and they propose effective pre-processing, data balancing and feature engineering techniques to get improved performance.

[34] applied part of speech tagging to identify suicide note classification. Their results show that the best ML classifier managed to attain a 78% classification rate out of the eight classifiers. However, their study used Natural Language Processing. [35] did a study where they predicted tuberculosis using LR and the DT classifier. They report an Accuracy of 98% on both classifiers and Area Under Curve (AUC) of 61.74% for LR and 59.28% for DT. Nevertheless, they used data from the Thai ministry of public health. [36] applied the one-against-all ensemble classifiers and NB to determine cause of death from VAs. Their approach showed improved model performance with a surge in Sensitivity scores of between 6% to 8%.

VA Data Challenges and Limitations

[1] argues that these VA algorithms and ML approaches can not avail enough evidence where there is limited expert diagnosis, hence they cannot be used to guide health priorities. These VA algorithms mainly employ statistical approaches and tariff scores to rank causes of death [15, 9]. Various VA algorithms and ML approaches are dependent on sample size, age group, causes of death, data set size or characteristics of the sample in order to produce best results [16, 4]. There has

been challenges and issues in terms of various VA techniques and there has been proposals to improve VA approaches through minimising the number of features under study (dimensionality reduction) and also combining various algorithms in order to improve on performance, accuracy and efficiency [13]. Moreover, the validity of the VA and ML approaches performance in terms of Sensitivity, Specificity and Predictive values vary with regard to causes of death across populations [37].

It is difficult to generalize and standardise VA classification practices, since there is no gold standard and patient's records vary in terms of socio-economic status [13, 15]. Another issue is associated with standardising VA questionnaires so that they have the same structure and content. The format and standard of VA questionnaires differs considerably, thus their administration requires appropriate training so as to elicit relevant and appropriate symptoms and causes. There are also often language barriers and the interviewer and interviewee need to speak the same language so as to derive best results. [4] recommended incorporating fully trained multiple translators. Additionally, the VA data collection procedures vary and is done by people who are not health professionals and possess different competencies. The other downside is the length of the recall period which can create a bias in the collected VA data. The heterogeneity of various autopsies in terms of the non-intersecting dialects of the English language (terms being in the native language) compromises data quality as most of these approaches tend to omit such autopsies in their model prediction, yet they might entail valuable information. This needs to be addressed to improve data quality of the VA narratives that are taken as input to the CCVA and ML approaches, thus eliminating possibilities of bias and mis-interpretations of the models.

This study seeks to close the gap in the existing body of knowledge by applying robust ML approaches for determining causes of death from VAs. This will be achieved by employing effective strategies for preprocessing, feature scaling, data balancing and feature engineering.

Methods

Study design

This is a retrospective cross-sectional study that uses secondary data analysis. All the cleaned VA narratives, model performance and classification results of various tasks are pushed from a Python Jupyter Notebook environment and housed within a PostgreSQL Version 4.2 object-relational database management system.

Population

This study uses only VA narratives data from the study area of the Agincourt Health and Demographic Surveillance System (HDSS). On a historical perspective the study setting was established in 1992 and is situated in the rural Sub-district of Bushbuckridge under Ehlanzeni District, in Mpumalanga Province, in north-eastern South Africa. The study area covers approximately 420km². According to Agincourt fact sheet of 2019, the population was at 116 247 individuals residing in 28 villages with 22 716 households, with males being 55 961, females being 60 280, children under 5 years being 11 724 and school going children with ages from 5 – 19 being 35 928 [38].

Data Source

Our dataset is from the Agincourt HDSS which is a surveillance site that specifically provides evidence based health monitoring that seeks to strengthen health priorities, practice and inform policy. The VA narratives data is for the period of 1993 to 2015. However, the doctors reviewed and classified cases are from 1993 to 2010 and suffice for our model training and prediction. We specifically use the VA narrative data and the corresponding cause of death assigned by two physicians and where they do not agree a third physician is consulted and assign a corresponding International Classification of Diseases-10 (ICD-10) code for each record in the dataset. Our data had 287 columns/features and 16338 records/observations. However, for this study we only used the VA narrative column (predictor = X), which where in English and in free text and the cause of death column assigned by certified physicians with a corresponding ICD-10 code on cause of death for each record. This implied now having only one narrative column and 16338 records. We further created twelve cause of death categories with corresponding number of samples for each class, as in Table 1. The cause of death categories where derived based on InterVA user guide and literature studies of [11, 27, 39, 3].

Table 1: Table showing disease categories, corresponding labels and number of samples per category

Category	Class Label	Number of samples
HIV/TB	0	3388
Other infectious	1	3388
Metabolic	2	3388
Cardiovascular	3	3388
Indeterminate	4	3388
Maternal and Neonatal	5	3388
Abdominal	6	3388
Neoplasms	7	3388
External causes	8	3388
Neurological	9	3388
Respiratory	10	3388
Other NCD	11	3388

Schematic Diagram of the Processes Followed

Figure 1 illustrates our logical steps that we follow for this study experiments. We first do data acquisition of our VA narratives as a comma separated value text file (csv), followed by data exploration and cleaning. Additionally, we do feature engineering, data balancing and feed our data to our models for training, validation and testing. Lastly we do cause of death classification.

[width=4cm]MLSchematicDiagram.JPG

Figure 1: High level Schematic Diagram of our ML process

Data Cleaning and Labelling

Data cleaning entails pre-processing were we aim at doing away with irrelevant data in order to improve model performance. After importing the dataset in csv format, we started by cleaning the unstructured narrative data by converting all text to lowercase, removed all punctuation, removed spaces, numbers and special characters.

Furthermore, we applied the TextBlob Python library to correct spellings. We performed tokenisation which is splitting a document (seen as a string) into tokens. All stop words were removed using the NLTK library of English stopwords removing insignificant words. Stemming was done to convert all possible word variations into the root form using the Python PorterStemmer library. Our dataset had certain missing values (narrative nulls where 2247 and cause of death nulls where 5170). We then dropped all nulls to remove bias in our modelling. Thereafter, feature engineering was done to determine the most representative features, as we then aimed at retaining only relevant words in the vector space by applying a weighting scheme [40]

Feature Engineering

Feature engineering was performed in order to generate new input features from existing ones. Feature engineering is made up of three steps namely, feature extraction, feature selection and feature value representation. We started by applying automated feature extraction as we aimed at only pulling out useful features using n-grams. n-grams are feature extraction techniques that identify feature(s) corresponding to token(s) where a token can be thought of as a string of words. n can be any number, $n = 1$ is a unigram, $n = 2$ is a bigram and $n = 3$ is a trigram [26]. Moreover, after the feature extraction we did feature value representation or term-weighting using the TF-IDF approach. Feature value representation is a process of creating a numeric vector of features, where each feature will have a corresponding numeric value that can be used for model learning. TF-IDF considers a feature important if the feature occurs frequently in the VA narratives belonging to one class and less frequently available in narratives belonging to another class. Feature selection was done to attain a the most relevant subset of features from the narratives using Singular Value Decomposition (SVD) as a selection criteria to reduce the dimension of our feature space. SVD creates a matrix in a low dimensional space and generates a matrix that is an exact representation of data. Moreover, it removes the less important terms producing an equivalent representation using any number of dimensions. This implies reducing our dataset containing a large number of values to a dataset containing significantly fewer values without loss of data [41, 42].

Data Balancing

Our dataset had data imbalances where the classes were imbalanced, meaning that there is a high difference between the positive values and negative values (e.g. more HIV positive than HIV negative classes). Furthermore, the (majority class:HIV/TB) had more samples as compared to other minority classes. This means that the minority classes were less represented in terms of data samples. As such, this creates bias in that minority classes as they will have fewer data points that can cause large misclassification errors. In order to address the issue of data imbalance we used the oversampling approach known as the Synthetic Minority Oversampling Technique (SMOTE). SMOTE was applied by generating artificial samples for the minority class, through interpolation between the positive instances that lie together. This approach addresses the issue of over-fitting caused by the general oversampling approach that replicates existing positive cases [8]. After data balancing we fed the data into our twelve models for training and validation.

Training, Validation and Testing

In this study, data was split into 70% training, 20% validation and 10% testing for all our twelve models. Training data is used to train our models and validation data is used to design and see how well our models perform so as to be able to check if our models are overfitting or underfitting. As such we can enforce some control mechanisms to address such. On the contrary, the test set is used to see how well a model performs with new or unseen data [42].

Machine learning models for classification

All our models were generated in Python using the Scikit learn module. Various supervised ML models were applied in this study to predict the ICD-10 related cause of death by taking input of the VA narratives and feeding into eight classifiers (SVM, DT, KNN, RF, Bagging, LR, NB and ANN). Our feature space was made up of inputs of VA narratives and our response variable was a categorical ICD-10 code for cause of death.

The NB is a statistical approach that uses conditional probabilities of each feature, thus assuming independence on predictors to assign cause of death to a record [26, 3]. This implies that all features or variables are independent, thus making it an effective classifier. This approach is also discussed in [41].

The LR is a statistical approach that uses maximum likelihood estimation to categorise classes. This approach seeks to predict a continuous or numeric variable by fitting a straight line or hyperplane to the feature space. It assumes a set of predictors or independent variables with corresponding categorical or response variable, such that it aims at to predict the probability of the response variable based on the independent variable. This makes logistic regression a classification approach. Similar approaches are used in the studies of [22, 41].

Bagging also known as bootstrap aggregation, is an approach that builds a classifier for every bootstrap based on a number of bootstraps from given datasets. The learning environment generates a classifier for every model from the dataset and a final classifier is created from aggregation of the classifiers that are classified from instances through voting for a class. It then chooses the class with most votes as the best classifier [43, 41].

DTs are tree based classifiers that split the feature space into rectangles using a technique known as recursive binary splitting. At first all features are within one rectangle and the process iterates through all dimensions and determines a split where there is a largest reduction in error measure (difference between the predictor and response is minimal). DTs define each node as a condition on a feature and a branch as a result of the condition and each leaf node as a class label. The start of the tree node is known as the root node (established through finding the feature that best divides the feature space), labelled as terms are branches with a weight. Leaf nodes represent class labels and new data points can be generated and assigned a class based on the majority vote of the leaf nodes [26]. Similar work is also reported in [41, 42].

The RF gets to a decision by using various decision trees to categorise new data points. Thus, it uses many individual learner trees to perform classification based on votes on an overall category from a given set of inputs. A majority vote is

then applied to determine a class of a new data point through combining random forest decision trees. It is similar to bagging except it builds trees based on random sampling of subset features at each node in the DT[33, 41, 42].

KNN uses a distance function to compute and generate new data points from available data points [3, 26]. This approach first creates a vector space for each data point in the training set, such that when a new data point is fed into the model for classification, this approach checks data points from the training set that are close to the new data point using the distance function. It then classifies this new label through combining the distance of the closest training examples [42].

SVM employs the idea of a decision boundary known as a hyperplane that distinguishes between classes in a high dimensional space. As such, it maximises the margin, which is the distance between the closest points of training set and hyperplane. The points that are close to the hyperplane are called support vectors. A classifier learns by maximising the margin between classes and uses the kernel function to map the narratives as input features to a high dimensional space. For cases that are not linearly separable, the SVM uses the kernel function [3, 26]. The SVM approach is also elaborated in the work of [42, 41, 44].

ANN are made up of a combination of perceptrons known as nodes. The output of a layer of nodes becomes input to the next layers. At the last stack of nodes is the output layer which generates the final output of the neural network. These ANN require vast amounts of training examples or data points [42]. This approach makes use of layers (input, hidden and output) that are a network of units where each unit can be a term and the output unit represents a category. In document text classification, weights are assigned to input units and the activation of these units is propagated through a network and the value of the output unit determines the categorisation. The layering of the nodes provides a map of the decision space also known as the neural network where a program can learn rules from massive data amounts being processed [45, 46, 40, 24]. We use a feed forward neural network that applies back propagation and data moves from one layer to the other layers using weights, biases and activation function to produce output. We further use the rectifier linear unit (reLu) activation function rather than tanh and sigmoid as they suffer from vanishing gradient (where the derivatives of the activation functions are closer to zero, thus smaller gradients mean the weights and biases will not be updated). The neural network function that we use is also reported in [41].

Model Optimisation

This study used 10 k-fold cross validation as an optimisation technique in order to evaluate our prediction models. This was done by first dividing our dataset into a training set, for model training and a test set for model evaluation. The dataset was randomly split into k (10) equal sized samples, where nine folds ($k - 1$) were used for model training and one fold was used as the holdout validation dataset for model testing. The cross validation process was iterated 10 times and at each iteration each of the k samples being used once for validation purposes. Furthermore, at each iteration an error is computed. The final result will be an average error of the model generated in each iteration. This implies that the k results from the folds are combined to produce a single estimation. The advantage of this approach is that, all

observations are used for both training and validation, with each observation used for validation exactly once. On the contrary, this approach has a disadvantage of having to define the number of folds manually. In order to address the limitations of the k-fold cross validation technique, we also used the automated GridSearch approach that eliminates the random setting of parameters and chooses optimum parameters automatically for a specific model. For tree based methods we applied cost complexity pruning (pre-pruning of trees) to get a smaller subset of trees that minimised our cost function by tuning parameter alpha through k-fold validation. We further set the minimum number of samples required at a leaf node and also set the maximum depth of the tree.

We employ the Mean Squared Error (MSE) and Cross Entropy Error (CEE) as cost functions for our ANN model. We further hyper-tune parameters such as the gradient descent, learning rate and back propagation to optimise our weights and biases on our network. This helps us achieve an optimal cost function that minimises the difference in predicted and response variables. We apply L_1 and L_2 regularisation approaches to further optimise our ANN model. L_1 regularisation involves eliminating features that are not useful for model prediction by setting some weights close to zero. On the contrary, L_2 regularisation tends to penalise large weights more and small weights less. Therefore, this approach is computationally efficient as an increase in regularisation tends to result in weights decaying towards zero. In this study, we follow the mathematical approach described in [42]. For most of our models we employ the MSE, Minkowski, Gini and CEE as cost functions to compute the minimal cost error between our predictor and the response using the k-fold cross validation approach to optimise model performance. These cost functions are described in [41].

Model Evaluation

Performance evaluation of classifier can be evaluated using various metrics and we report the metrics based on studies by [33, 26]. We present Accuracy, Precision, Recall, F-score and Area Under Curve (AUC) as our metrics for evaluation. In order to compute these metrics we use the following values from the confusion matrix; True Positives (TP) denoting predicted positive VA narratives with a particular disease category from the twelve classes and are actually positive. False Positives (FP) predicted positive VA narratives with a particular disease category from the twelve classes but are actually negative. True Negatives (TN) denoting predicted negative VA narratives with a particular disease category from the twelve classes and are actually negative. False Negatives (FN) implying the predicted negative VA narratives with a particular disease category from the twelve classes but are actually positive. Precision also known as the Positive Predictive Value (PPV) defines the proportion of VA narratives correctly predicted as positive to the total of positively predicted VA narratives. Recall also known as Sensitivity or True Positive Rate (TPR) defines the proportion of VA narrative correctly predicted as positive to all VA narratives in the actual positive category. F-measure computes the average or harmonic mean of Precision and Recall. Accuracy denotes all classes with classified results that have been predicted correctly in fraction terms. The Receiver Operating Characteristic Curve (ROC curve) visualises the TPR against the false positive rate

(FPR). Area under the ROC curve applies the principle of plotting a curve specific to a machine learning algorithm where the classifier is evaluated relative to a weighting on the area under the curve. Good performance of the algorithm is given a weight of close to 1, thus graph is AUC closer to upper left corner and the poor performance of an algorithm is given a weight of 0.5 and below. Specificity computes the ratio of negative VA narratives that are correctly predicted as negative.

Results

In this section we present the results attained from various classification techniques employed to determine cause of death from VA narratives.

Performance evaluation of ML classifiers

Precision, Recall, Accuracy and F-score measure for our eight classifiers in the cause of death categorisation of the twelve disease classes are presented in Table 2. We also present the ROCAUC curve for our twelve disease categories in Figure 2.

The RF classifier outperformed all the other eight classifiers with a Precision of 95%, Recall of 95%, F1-score of 95% and Accuracy of 95%. The second best performing classifier was the ANN with a Precision of 94%, Recall of 94%, F1-score of 94% and Accuracy of 94%. The third best classifier in terms of performance was KNN with a Precision of 93%, Recall of 93%, F1-score of 92% and Accuracy of 93%. The SVM was the fourth best performing and achieved a Precision of 92%, Recall of 92%, F1-score of 92% and Accuracy of 92%. We applied bagging as our ensemble classifier and attained a Precision of 91%, Recall of 91%, F1-score of 91% and Accuracy of 91%. Our sixth classifier was the DT which attained a Precision of 84%, Recall of 85%, F1-score of 84% and Accuracy of 85%. The LR attained a Precision of 82%, Recall of 82%, F1-score of 82% and Accuracy of 82%. The least performing classifier was the NB with a Precision of 75%, Recall of 71%, F1-score of 72% and Accuracy of 71%. Table 2 shows the performance evaluation of our eight models.

The results of the ROCAUC show that our models were good classifiers of cause of death categorisation of the twelve disease categories. The classes 3,6,9,10 and 12 have an area under the ROC curve of equal to 1. They are followed by classes 8, 2 and 7 with an area under ROC curve of 0.99, 0.97 and 0.95 respectively. The other classes have an area ROC curve within the range 0.84 – 0.89. We report a micro-average ROC curve AUC of 0.96 and a macro-average ROC curve AUC of 0.95.

Table 2: Performance evaluation of models

Model name	Accuracy	Precision	Recall	F1-score
RF	95%	95%	95%	95%
ANN	94%	94%	94%	94%
KNN	93%	93%	93%	92%
SVM	92%	92%	92%	92%
Bagging	91%	91%	91%	91%
DT	85%	84%	85%	84%
LR	82%	82%	82%	82%
NB	71%	75%	71%	72%

All list of abbreviations and corresponding terms are given in Table 3.

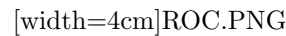
A placeholder for a Receiver Operating Characteristic (ROC) curve image, labeled as ROC.PNG.

Figure 2: Performance evaluation using the ROCAUC

Discussion

Determining cause of death from VAs largely remains a manual task that is tedious, time consuming, prone to errors and costly. Attaining quality narratives from the interview process requires trained interviewers that can elicit valuable information from the interviews. Despite efforts to improve the VA processes, the VA elicitation process suffers from many drawbacks and still lags behind in the determination of cause of death from VAs. This ultimately affects VA reporting and does not happen in real time, even though they are key in informing civil registration systems and strengthening of health priorities. Despite most studies in VA using responses from the standardised questionnaire for ML prediction, this study synthesised and analysed the robustness of ML techniques in determining causes of death from VA narratives only. Our findings suggest that ML has already been applied in the VA domain. Nevertheless, strategies of text pre-processing and handling of data are not standardised and are inconsistent, hence this affects model performance. Existing literature lacks clear explanations on how data exploration is done and strategies implemented post data exploration.

To support our findings, this study applies effective data exploration, pre-processing, feature scaling and data balancing. There is a lack of studies that report on the application of effective pre-processing techniques, a step crucial in the data preparation for optimum model performance. Most work done in ML does not report or apply any data balancing techniques. As such, this leads to bias and false interpretation of results. There is need to handle data imbalance prior to building machine learning models for training as this affects model performance. As such, implementing proper data balancing enforces transparency and conforms to the standards of explainable AI. Most studies in VA that apply feature engineering only do feature extraction and feature value representation. Little research reports on feature selection being applied in VA data [8]. As such, most of these datasets are modelled with many noisy data points. There is a great need to do feature selection to curb over-fitting and reduce the highly dimensional feature space and only retain relevant features for our models. Because of less application of feature selection, we continue to witness low performing ML models in the VA domain. Improving model performance requires application of optimisation techniques. Most studies in the VA domain only use k-fold cross validation as an optimisation approach. However, there is a new novel automated optimisation technique known as GridSearch. Unlike k-fold cross validation, where one has to manually pre-define the number of k folds, the GridSearch automates the process of hyper-tuning of parameters leading to finding the optimal alpha that can be used for model fitting. In this study we use cost functions as penalties to attain the smallest error distance between the response and predictors. This process helps us achieve optimum results of our models.

It is imperative for researchers who intend to use ML approaches to strike a balance between interpretability and accuracy. Our results, consistent with a number

of studies that used VA data to determine cause of death, suggest that ML approaches can accurately classify cause of death from VA narratives. However, in most cases statistical approaches are always outperformed by other ML approaches [3, 33, 28, 26, 8, 13, 9]. After applying effective data handling strategies, we discuss on the performance of our eight classifiers in subsequent paragraphs.

The RF classifier achieved the best results as compared to the other classifiers. This can be attributed to effective data pre-processing, cleaning and feature selection strategies that enormously reduced the feature space leaving only discriminative informative features for the model. This suggests an effective generation of a forest with powerful low depth trees that gave us good results and model performance. This approach can handle both categorical and numeric features and produces an easily interpretable model. Second on the list in terms of performance was the ANN which is capable of learning complex patterns and relationships within data. It has effective preprocessing that is inbuilt within the abstract layers. Additionally, employing back propagation, use of weights, bias and learning rate through application of an effective activation function suggest the good results attained by this model. This suggests that with more layers this model can improve in terms of performance. Consequently, this calls for an urgent need to explore with deep learning architectures. KNN was our third best performing classifier and it calculated the similarity between new VA narrative data and the VA narratives for the training data finding the number of k most similar cases which are then labelled as new class instances based on extracted majority VA narratives. The results suggest this model computed the distance measures effectively and accurately. This suggests that we did optimum linear scaling of features which were discriminative. Moreover, we managed to apply effective dimensionality reduction schemes. Additionally, this model is intuitive and thus it is a training set itself. This approach is effective with numerical data and multidimensional data. However, using this technique with categorical features and limited number of values can still yield better results. Therefore, getting best results depends on the parameters that you select such as the number of neighbours, distance measure and kernel function to use.

On the contrary the SVM performs better than Bagging our ensemble classifier, DT and our statistical models (NB,LR). This implies that the categorisation task was not linearly separable as it uses margins to perform categorisation. Moreover, the SVM approach is immune to over-fitting and high dimensionality of feature space because its feature space assumes independent features. Additionally, this model performs better when faced with high dimensional data and numerical features. However, it faces challenges of large memory requirements, model complexity and interpretability [8]. Our ensemble approach bagging produced better results as compared to DT, LR and NB but a bit lower to SVM, KNN, ANN and RF. This can be attributed to the fact that the ensemble classifier combines all capabilities of all classifiers including strengths and weaknesses and attains optimum results through voting. Even though the DT was outperformed by RF, ANN, KNN, SVM and bagging, its performance was moderate and managed to perform better than LR and NB. This can be attributed to the fact that a DT is not affected by data standardisation and it performs better at multi-class classification problems. Moreover, it can handle both continuous and categorical variables using more comprehensive

rules. The DT classifier can nonetheless suffer from over-fitting leading to complex trees that may need to be pruned and may require applying effective hyper-parameter tuning to avoid such. However, using a single DT model means a weak learner, thus having a multitudes of trees can produce better predictions as evidenced by the RF classifier in this study. Statistical models (LR and NB classifiers) and DT did not yield optimal results. The NB classifier performance highlights that the assumption of conditional independence among features might have been a key factor in poor performance. This implies that as the feature space increases, feature dependence creates complexity in the model performance. Additionally, this can be attributed to the fact that these approaches fail to handle non-linearity and fail to handle complex patterns in the dataset.

One key finding is that, VA researchers who intend to investigate performance of ML algorithms on VA data should aim to employ various statistical and ML algorithms to fully understand how they perform. Most studies have only applied own classifiers on their own datasets, and this lacks generalisation and comparability. Therefore, one cannot deduce that one classifier is best as the performance varies from one domain to the other. There is a need to explore with various ML approaches for comparative reasons and generalisations.

This study had limitations of the data quality which was incomplete and collected using various tools. These problems emanate from the issues of VA data being collected by non-medical personnel with different competencies and qualifications. Also the length of the recall period creates room for bias as narrations might be wrongly interpreted. Moreover, these narratives are given as summaries which might also not be a true reflection of events that transpired. Furthermore, there is a challenge of some of the VA narratives being in the native language or a combination of native and common English language. The capturers of these narratives also generate errors when transcribing the narratives. This calls for novel approaches of addressing such issues. One solution can be the use of trained personnel who can also work with qualified translators so as to elicit appropriate information. Another solution can be to use NLP to translate the vernacular terms into a common English vocabulary which machines can understand and process effectively. Ultimately this can improve model performance.

Conclusion

We successfully assessed the robustness of various ML approaches in determining cause of death from VA narratives only that proved to have rich valuable information. We specifically explored with eight ML algorithms and the RF, ANN, KNN, SVM and bagging in that order outperformed the other classifiers. Therefore, this further reinforces the notion that ML approaches can be used to determine cause of death from the VA narratives in real time, in a cost effective way that is free from human error thus also saving time. This study was limited in that it only used VA narratives from Agincourt HDSS with only twelve disease categories. The results of this study avail interesting opportunities to investigate ML determination of cause of death from VA narratives using large datasets from other sources with more cause of death categories. Moreover, we aim at exploring with deep learning architectures which have shown promising results in other domains and have not been fully exploited within the VA domain. Additionally, we aim at investigating ML model

performance using a combination of responses from the structured questionnaire and the VA narratives as they have proved to have valuable rich information. This will contextualise and reinforce added value in improving ML model performance, thus attaining more accurate results in cause of death determination from VAs.

Abbreviations

Table 3 is a list of abbreviations and corresponding full terms.

Table 3: List of abbreviations

Abbreviation	Full Term
ANN	Artificial Neural Network
AUC	Area Under Curve
AUCROC	Area Under Curve Receiver Operating Characteristics
CCVA	Computer Coded Verbal Autopsy
DT	Decision Tree
ICD-10	International Classification of Diseases-10
HDSS	Health and Demographic Surveillance Site
KNN	K-Nearest Neighbour
LMIC	Low to Medium Income Country
LR	Logistic Regression
ML	Machine Learning
NB	Naive Bayes
PCVA	Physician Coded Verbal Autopsy
PHMRC	Population Health Metrics Research Consortium
RF	Random Forest
SMOTE	Synthetic Minority Oversampling Technique
SVM	Support Vector Machine
TF	Term Frequency
TF-IDF	Term Frequency with Inverse Document Frequency
VA	Verbal Autopsy

Acknowledgements

This work was supported by the Developing Excellence in Leadership, Training and Science (DELTAS) Africa Initiative Sub-Saharan Africa Consortium for Advanced Biostatistics (SSACAB) [Grant No.DEL-15-005]. The DELTAS Africa Initiative is an independent funding scheme of the African Academy of Sciences (AAS) Alliance for Accelerating Excellence in Science in Africa (AESA) and is supported by the New Partnership for Africa's Development Planning and Coordinating Agency (NEPAD Agency) with funding from the Wellcome Trust [Grant No. 107754/Z/15/Z] and the United Kingdom government. Furthermore, we received support from the University of the Witwatersrand, Faculty of Health Sciences Seed funding.

Funding

This research study was not funded.

Availability of data and materials

The datasets generated and/or analysed during the current study are not publicly available [The Authors do not have permission to share the data used for this study] but are available from the corresponding author on reasonable request.

Ethics approval and consent to participate

This study has been approved by the relevant ethics committee in South Africa (University of the Witwatersrand Faculty of Health Sciences, Human Research Ethics Committee (Medical), approval ref. no: M1911132). We got permission from the Agincourt Health and Demographic Surveillance Site to use their data for research purposes.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not Applicable.

Authors' contributions

All authors contributed to study conceptualisation. MM wrote the original draft of the study protocol. MM, CK, EM and TC drafted and designed the objectives of the study protocol. MM, CK and EM ensured ethical approval of the study. MM, CK, TC and EM contributed to background and study design, focusing on clinical objectives. MM, TC and VO did the algorithm experiments and designed the models used in this study. All authors contributed to refine and finalize the study protocol. All authors read and approved the final manuscript.

Author details

¹ School of Public Health, Department of Epidemiology and Biostatistics, University of The Witwatersrand, Johannesburg, South Africa. ² MRC/Wits Rural Public Health and Health Transitions Research Unit (Agincourt), University of The Witwatersrand, Johannesburg, South Africa. ³ National Health Laboratory Service (NHLS), National Cancer Registry, Johannesburg, South Africa. ⁴ Wits Institute of Data Science, University of The Witwatersrand, Johannesburg, South Africa. ⁵ School of Electrical and Information Engineering, University of The Witwatersrand, Johannesburg, South Africa.

References

- Nichols EK, Byass P, Chandramohan D, Clark SJ, Flaxman AD, Jakob R, et al. The WHO 2016 verbal autopsy instrument: an international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. *PLoS medicine*. 2018;15(1):e1002486.
- Thomas LM, D'Ambruoso L, Balabanova D. Verbal autopsy in health policy and systems: a literature review. *BMJ global health*. 2018;3(2):e000639.
- Jebblee S, Gomes M, Jha P, Rudzicz F, Hirst G. Automatically determining cause of death from verbal autopsy narratives. *BMC medical informatics and decision making*. 2019;19(1):127.
- Soleman N, Chandramohan D, Shibuya K. Verbal autopsy: current practices and challenges. *Bulletin of the World Health Organization*. 2006;84:239–245.
- Lozano R, Lopez AD, Atkinson C, Naghavi M, Flaxman AD, Murray CJ. Performance of physician-certified verbal autopsies: multisite validation study using clinical diagnostic gold standards. *Population Health Metrics*. 2011;9(1):1–13.
- Mapundu MT, Kabudula C, Musenge E, Celik T. Overview of Statistical and Machine Learning Techniques for Determining Causes of Death from Verbal Autopsies: A Systematic Literature Review. 2020;.
- Reeves BC, Quigley M. A review of data-derived methods for assigning causes of death from verbal autopsy data. *International journal of epidemiology*. 1997;26(5):1080–1089.
- Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical text classification research trends: Systematic literature review and open issues. *Expert systems with applications*. 2019;116:494–520.
- Desai N, Aleksandrowicz L, Miasnikof P, Lu Y, Leitao J, Byass P, et al. Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low-and middle-income countries. *BMC medicine*. 2014;12(1):20.
- James SL, Flaxman AD, Murray CJ. Performance of the Tariff Method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*. 2011;9(1):31.
- Byass P, Herbst K, Fottrell E, Ali MM, Odhiambo F, Amek N, et al. Comparing verbal autopsy cause of death findings as determined by physician coding and probabilistic modelling: a public health analysis of 54 000 deaths in Africa and Asia. *Journal of global health*. 2015;5(1).
- McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*. 2016;111(515):1036–1049.
- Miasnikof P, Giannakeas V, Gomes M, Aleksandrowicz L, Shestopaloff AY, Alam D, et al. Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC medicine*. 2015;13(1):286.
- Clark SJ, Li Z, McCormick TH. Quantifying the Contributions of Training Data and Algorithm Logic to the Performance of Automated Cause-assignment Algorithms for Verbal Autopsy. *arXiv preprint arXiv:180307141*. 2018;.
- Clark SJ. A Guide to Comparing the Performance of VA Algorithms. *arXiv preprint arXiv:180207807*. 2018;.
- Kalter HD, Perin J, Black RE. Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death. *Journal of global health*. 2016;6(1).
- Quigley MA, Chandramohan D, Setel P, Binka F, Rodrigues LC. Validity of data-derived algorithms for ascertaining causes of adult death in two African sites using verbal autopsy. *Tropical Medicine & International Health*. 2000;5(1):33–39.
- Murray CJ, Lozano R, Flaxman AD, Serina P, Phillips D, Stewart A, et al. Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC medicine*. 2014;12(1):1–19.
- Leitao J, Desai N, Aleksandrowicz L, Byass P, Miasnikof P, Tollman S, et al. Comparison of physician-certified verbal autopsy with computer-coded verbal autopsy for cause of death assignment in hospitalized patients in low-and middle-income countries: systematic review. *BMC medicine*. 2014;12(1):22.
- Nithya B, Ilango V. Predictive analytics in health care using machine learning tools and techniques. In: 2017 International Conference on Intelligent Computing and Control Systems (ICICCS). IEEE; 2017. p. 492–499.
- Flaxman AD, Vos T. Machine learning in population health: Opportunities and threats. *PLoS medicine*. 2018;15(11).
- Mwanyangala MA, Urassa HM, Rutashobya JC, Mahutanga CC, Lutambi AM, Maliti DV, et al. Verbal autopsy completion rate and factors associated with undetermined cause of death in a rural resource-poor setting of Tanzania. *Population health metrics*. 2011;9(1):41.
- Quigley MA, Chandramohan D, Rodrigues LC. Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. *International Journal of Epidemiology*. 1999;28(6):1081–1087.
- Bouille A, Chandramohan D, Weller P. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. *International journal of epidemiology*. 2001;30(3):515–520.
- Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population health metrics*. 2011;9(1):29.
- Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K. Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. *Journal of forensic and legal medicine*. 2018;57:41–50.
- Danso S, Atwell E, Johnson O. A comparative study of machine learning methods for verbal autopsy text classification. *arXiv preprint arXiv:14024380*. 2014;.

28. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Classification of forensic autopsy reports through conceptual graph-based document representation model. *Journal of biomedical informatics*. 2018;82:88–105.
29. Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC medical informatics and decision making*. 2015;15(1):53.
30. Danso S, Johnson O, Ten Asbroek A, Soromekun S, Edmond K, Hurt C, et al. A semantically annotated Verbal Autopsy corpus for automatic analysis of cause of death. *ICAME Journal*. 2013;37.
31. Danso S, Atwell E, Johnson O, ten Asbroek G, Edmond K, Hurt C, et al. A verbal autopsy corpus for machine learning of cause of death. In: *Proceedings of the Corpus Linguistics Conference*; 2011. .
32. Khoozani ZS, Raj RG. TF-IDF-Based Automated Application for classification Forensic Autopsy Reports to Identification of Cause of Death (CoD);.
33. Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PLoS one*. 2017;12(2):e0170242.
34. Pestian J, Nasrallah H, Matykiewicz P, Bennett A, Leenaars A. Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*. 2010;3:BI1–S4706.
35. Taufik MR, Lim A, Tongkumchun P, Dureh N. Predicting TB Death Using Logistic Regression and Decision Tree on VA Data;.
36. Murtaza SS, Kolpak P, Bener A, Jha P. Automated verbal autopsy classification: using one-against-all ensemble method and Naïve Bayes classifier. *Gates open research*. 2018;2.
37. Chandramohan D, Setel P, Quigley M. Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *International Journal of Epidemiology*. 2001;30(3):509–514.
38. Kabudula CW, Tollman S, Mee P, Ngobeni S, Silaule B, Gómez-Olivé FX, et al. Two decades of mortality change in rural northeast South Africa. *Global health action*. 2014;7(1):25596.
39. King G, Lu Y, et al. Verbal autopsy methods with multiple causes of death. *Statistical Science*. 2008;23(1):78–91.
40. Korde V, Mahender CN. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*. 2012;3(2):85.
41. Zaki MJ, Meira Jr W. *Data Mining and Machine Learning: Fundamental Concepts and Algorithms*. Cambridge University Press; 2019.
42. Leskovec J, Rajaraman A, Ullman JD. *Mining of massive data sets*. Cambridge university press; 2020.
43. Fawagreh K, Gaber MM, Elyan E. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*. 2014;2(1):602–609.
44. Poole DL, Mackworth AK. *Artificial Intelligence: foundations of computational agents*. Cambridge University Press; 2010.
45. Byrne MD. Machine Learning in Health Care. *Journal of PeriAnesthesia Nursing*. 2017;32(5):494–496.
46. Iqbal Z, Ilyas R, Shahzad W, Inayat I. A comparative study of machine learning techniques used in non-clinical systems for continuous healthcare of independent livings. In: *2018 IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE)*. IEEE; 2018. .

Figures

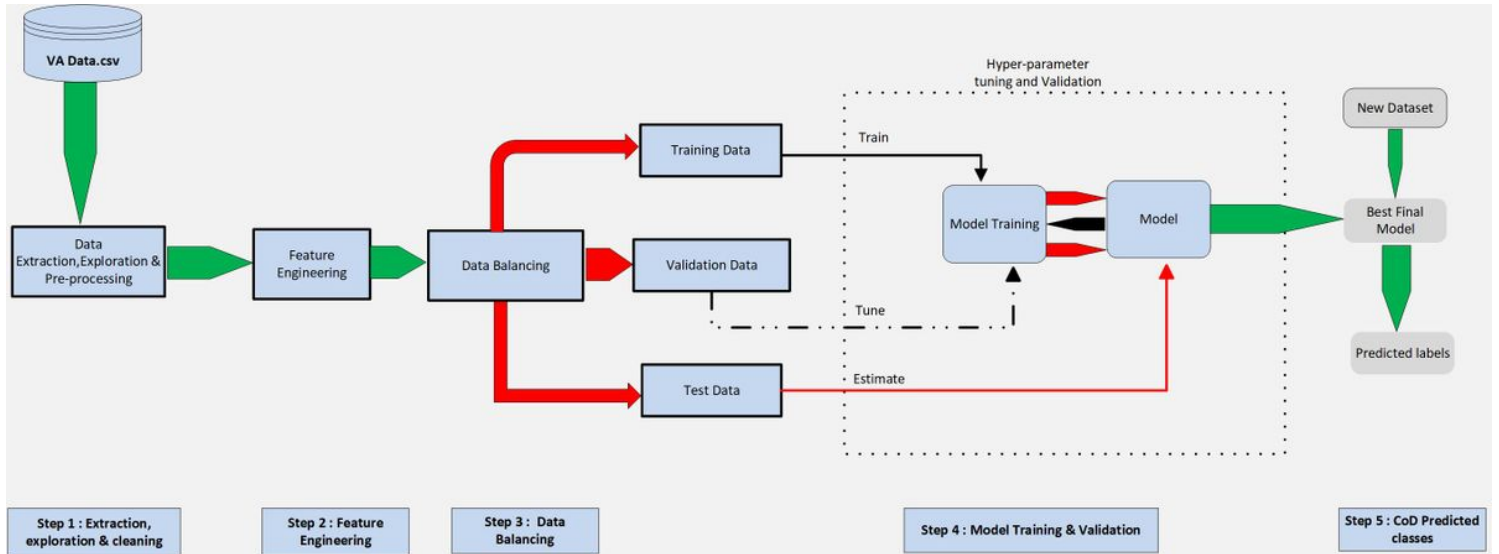


Figure 1

High level Schematic Diagram of our ML process

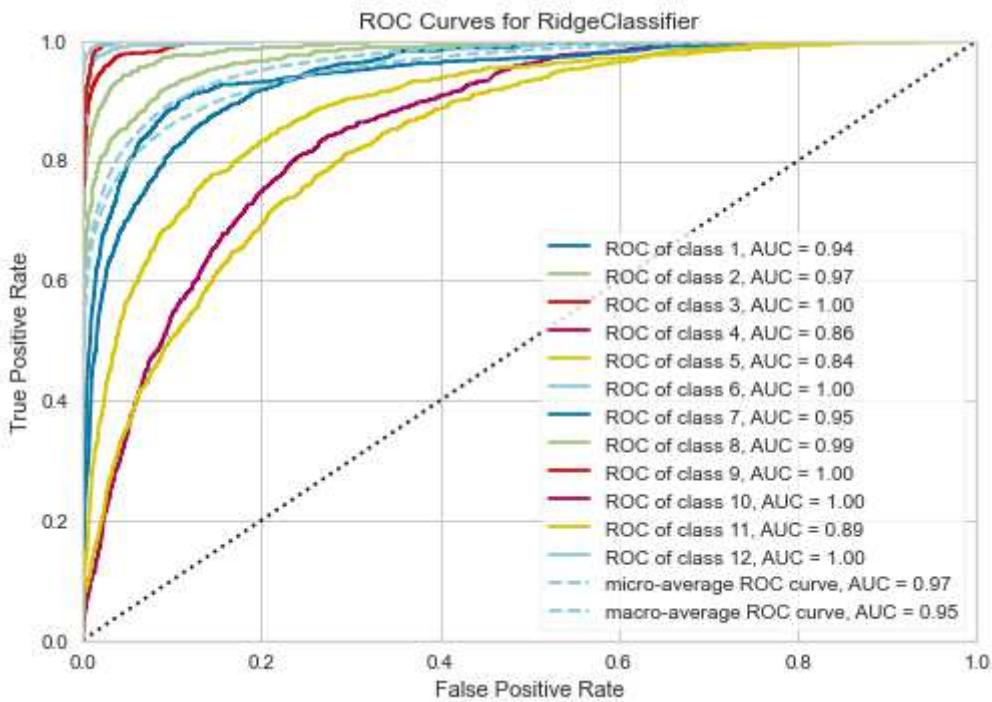


Figure 2

Performance evaluation using the ROCAUC