A Machine Learning Framework and Case Study for Exploring Mortality in Developing Countries with Verbal Autopsies

INTERNSHIP REPORT

Submitted by
ANDY LEE

In partial fulfillment of the requirements
for the degree of
Master of Science in Applied Computing (MScAC)
Department of Computer Science, University of Toronto

Under the Guidance of
Industry Supervisor:
Dr. Richard Wen, PhD
Research Data Scientist, Centre for Global Health Research (CGHR)

Academic Supervisor:
Dr. Frank Rudzicz
Associate Professor, Department of Computer Science, University of Toronto
Associate Professor, Faculty of Computer Science, Dalhousie University

DECEMBER 2023

ABSTRACT

Verbal autopsy (VA) is a practice used to collect medical information about individuals shortly before their death. In situations where conventional autopsies are not performed, VA provides valuable information that physicians can use to determine the likely cause of death (CoD). While beneficial, the process is not optimal in terms of medical resource efficiency and CoD assignment reliability. Several VA tools utilizing statistics, probabilities, and expert algorithms have been developed for CoD assignments. However, research into using machine learning (ML) for this purpose has been inadequate and lagging. This report is divided into two parts: The first outlines a six-component conceptual framework developed for implementing ML in VA, from data collection to application. The second part presents a case study that applies this framework, using two rounds of data from Healthy Sierra Leone (HEAL-SL) to train models and predict CoD. The partial chance-corrected concordance (PCCC) scores for trained models range from 0.60 to 0.63, with predicted CoD PCCC scores between 0.55 and 0.62, where 0 indicates poor performance and 1 indicates perfect concordance. Furthermore, the logistic regression model yielded the highest average PCCC and F1-scores. These results suggest that ML methods can be a viable alternative for predicting and informing CoD assignments.

ACKNOWLEDGEMENTS

I extend my sincere gratitude to my supervisors, Dr. Richard Wen, Research Data Scientist at the Centre for Global Health Research (CGHR), St. Michael's Hospital, Unity Health Toronto, and Dr. Frank Rudzicz, Associate Professor at the Department of Computer Science, University of Toronto and the Faculty of Computer Science, Dalhousie University. Their invaluable guidance and support throughout my research have been instrumental. Their combined expertise in machine learning, natural language processing, and applied healthcare has profoundly shaped my work. I am also thankful to CGHR and Dr. Prabhat Jha, Director of CGHR, for providing students with the enriching opportunity to contribute to improving public health research.

I appreciate the unwavering support of my family and friends, whose encouragement has been a constant source of motivation.

Special thanks to Mitacs for their financial support in facilitating my research.

Introduction

1. Background

The internship was conducted at the Centre for Global Health Research (CGHR), a non-profit organization sponsored by St. Michael's Hospital, Unity Health Toronto, and the University of Toronto. The primary function of CGHR is to conduct large-scale public health and epidemiology studies that impact public health systems and policies, particularly in low- to middle-income countries (LMICs).

One of the many functions of CGHR is to research the utility of verbal autopsy (VA) in LMICs. VA is a method that records events preceding a person's death through interviews with family or friends [1], [2]. Unlike conventional autopsy, VA does not involve a physical examination of the deceased's body by a medical professional. Traditionally, physicians review each VA record to determine the cause of death (CoD), a process known as Physician-Coded Verbal Autopsy (PC-VA). This process has several drawbacks and considerations. For example, it is inherently resource-intensive, and some argue that it diverts physicians' time from treating the living [3]. Furthermore, the ability of physicians or VA tools to maintain consistent CoD assignments is also a matter of frequent concern [4], [5]. These reasons create an incentive to automate the CoD prediction process. Over the past decade, various automated tools, sometimes referred to as Computer-Coded Verbal Autopsy (CC-VA) or automated methods, have been developed, mainly utilizing domain expertise [1], [6]–[10] and probabilistic methods [1], [8], [9], [11]–[16] to determine CoD.

As such, CGHR aims to achieve two objectives through the internship program:

The first objective is to leverage the recent advancements in machine learning (ML) to investigate and build a conceptual framework on how VA data can be used with ML, and to identify the potential applications for these ML models. The goal of the conceptual framework is to help public health professionals who heard of, but are not familiar with,

ML tools and methods to understand the components involved in implementing the entire process from VA data to application.

The second objective is to conduct a case study using the HEAL-SL dataset [17] to predict CoD. The goal is to put the conceptual framework into practice by applying it in a case study.

2. Literature Review

To create the conceptual framework, we first need to assess the current landscape of how verbal autopsies are used in conjunction with automated methods. A scoping systematic review approach was employed, which involves the intern conducting the review process following the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) framework. The search employed the following keywords: verbal autopsy, machine learning, clustering, artificial intelligence, natural language processing, automation, computer coded, computer-coded, prediction, and automated method. Databases searched include OVID, Pubmed, IEEE, Scopus, Web of Science, focusing on papers published since 2013.

From these sources, 609 studies were screened, and 261 duplicates were removed. 348 studies were screened for title and abstract, with 235 studies deemed irrelevant, leaving 113 studies for full-text assessment. The following keywords were highlighted to help identify relevancy: InterVA-4, InterVA-5, SmartVA, InSilicoVA Bayes, Tariff, Classification, Transfer, Bagging, Clustering, neural network, regression, physician, computer-coded, PCVA, CCVA, population level, individual level, machine learning, artificial intelligence, mortality, PHMRC, natural language, feature, performance, verbal autopsy, verbal autopsies, and invasive tissue. Any abstract that did not mention statistical method or machine learning were excluded.

The remaining literature underwent a full-text assessment. Key takeaways were extracted and categorized based on its relevancy to the VA process, assignment methods, applications, improvements, model performance studies, real-world application, etc. Additionally, relevant topics such as providing context to the origins of a particular automated method were included manually on an individual basis.

3. Creating the Conceptual Framework

Based on the knowledge gathered from the literature review, we constructed the conceptual framework featuring six major components and how they are interconnected.

3.1 Verbal Autopsy Data

The first component is “VA data”, which consists of three parts: structured data from the questionnaire, typically in formats more suitable for processing by statistical models; open narrative, an unstructured text commonly utilized by physicians; and physician keywords, key points extracted by physicians based on their assessment using both the questionnaire and the open narrative.

3.2 Preprocessing

The second component, “Preprocessing”, lists common preprocessing techniques used to process raw VA data. While these techniques are optional, some are considered best practices aimed at enhancing data quality, which translates to better model performance [18]–[20]. Standard preprocessing primarily refines structured data by identifying and correcting errors, removing redundancies, managing missing information, and converting data into appropriate formats for further analysis. After preprocessing, the subsequent “feature engineering” involves creating, modifying, or selecting relevant features to improve the performance of models. Natural Language Processing (NLP) is another field aimed at processing and transforming unstructured text into suitable

formats [10], [21]–[23]. It extracts meaningful patterns and relationships from data that can be used in the models, mainly for open narratives and physician keywords.

3.3 Modelling

Third component in “Modelling” consists of models used to perform analysis on the VA data, some of which are specifically created for this purpose. Statistical and probabilistic models use statistical techniques and probability distributions to assign CoD, such as InterVA, InSilicoVA [5]. ML models are more general-purpose computer algorithms that leverage patterns in data to “learn” and iteratively improve model performance, such as Logistic Regression and k-Nearest Neighbour [24]. Deep learning models are advanced ML models that utilize vast amounts of data through neural architectures to identify relationships and patterns, such as Artificial Neural Network and Large Language Models [24], [25].

3.4 Model Optimization

Fourth component in “Model Optimization” include techniques from various disciplines to enhance model performance. For example, hyperparameter tuning involves testing and optimizing individual parameters for optimal results, while transfer learning addresses the challenges where models trained using data in specific medical settings, geographic locations, or sociocultural contexts, are tasked with making inferences in different contexts [26]–[29].

3.5 Evaluation

The fifth component, “Evaluation”, involves using common statistical metrics and domain-specific metrics to measure and compare CoD prediction performance between various models and physician assignments. These metrics include calculating individual-level prediction accuracies such as Partial chance-corrected concordance (PCCC), or population causes of death distributions such as Cause-specific mortality fraction (CSMF) accuracy [30].

3.6 Application

Finally, the sixth component, “Application”, showcases potential applications of VA. One of the most prominent issues in LMICs is the lack of systematic and organized documentation of diseases within a civil registration and vital statistics (CRVS) system to develop robust health policies. Implementing and maturing such systems can be time-consuming and resource-intensive, a challenge even for developed countries [1], [31], [32]. LMICs often face additional difficulties due to resource constraints. Consequently, the VA process, particularly the developments in automated methods, have emerged as potential solutions or interim approaches. VA using automated methods can be employed to study, validate, compare, or update data pertaining to specific burdens of diseases or injuries within or among certain demographics [31]–[44].

4. Case Study

The second objective of the internship involves a case study using HEAL-SL VA data to train various ML models to predict CoD. The case study constructs an ML pipeline that follows the conceptual framework developed earlier. The predictions are analyzed both at the individual and population levels, and the findings are presented to CGHR. The serves to enhance the understanding of using automated methods in VA and to support and inspire other current or future initiatives.

In addition to serving as a practical application of the conceptual framework, the case study aims to answer several research questions: Which ML model predicts the CoD best? How do the different components in the VA dataset contribute to CoD prediction? Which CoD can be best predicted by ML models?

4.1 Dataset

The datasets used and trained in this case study are the HEAL-SL datasets supplied by openmortality.org [45]. As of this writing, two rounds of data have been published, with the third round expected to be released in 2024. The data collection is a joint effort by the Ministry of Health and Sanitation of Sierra Leone (MOHS), Njala University, and

CGHR. This study will serve as a pivoting point for improving public health and overall CRVS system in Sierra Leone.

The two rounds of data contain VA information for over 11,000 deaths, from which we will only be using the adult age groups which is 15 years old and above. In both rounds, the sex distribution is approximately similar, with a slightly higher percentage of males than females. The Round 1 dataset consists of 55% male and 45% female, while the Round 2 dataset comprises 56% male and 44% female

The datasets are categorized into three data types: numerical for continuous values, categorical for discrete values, and textual for open text. Both rounds exhibit approximately similar percentage of each data type. Categorical data occupied 75% of the data, textual data occupied 15%, and the remainder is numerical.

The CoDs are coded based on the World Health Organization (WHO) International Classification of Diseases 10th Revision (ICD-10) [46], and each record can have up to five CoDs. Two physicians independently evaluate and compare the CoD in the “initial round”. If the codes match, the CoD is finalized. Conversely, if they do not match, the physician tries again in a “reconciliation round”. If there is still no consensus on the CoD, an adjudicator is brought in to determine the final CoD.

In this study, CGHR-10 titles [47] are used as target labels, representing CoDs in plain text. To convert ICD-10 codes into CGHR titles, we first obtain the last CoD determination, then map each ICD-10 code to its respective CGHR-10 title. This process will be explained further in the methodology section. The converted labels results to 18 unique CoD for round 1 dataset, which includes Malaria, Infections, Non-Communicable Diseases (NCDs), and so on. the round 2 dataset has 19 titles, with the additional CoD “Nutritional” not present in the previous round.

Appendix B – Cause of Death Distribution in HEAL-SL Datasets presents CoD distribution graphs that aggregate the CoD titles sorted in descending order to showcase the difference in distribution composition of CoD for both datasets. “Malaria”,

“Infection”, and “NCDs” have the highest percentage of records in both rounds but are ranked differently. The same applies to “Other injuries”, “Stroke”, and “Diarrhoeal” as the second highest percentage of records but ranked differently. The percentages of “Maternal” and “Cancers” records were almost halved in round 2 dataset. “TB”, “Liver & Alcohol”, and “Cardiovascular” saw an increase of about 1.5% in records. “Diabetes”, “Suicide”, and “Nutritional” remains lowest for both rounds with less than 1% of records.

4.2 Methodology

4.2.1 Overview

The round 1 dataset, which has approximately 5003 records, will be used exclusively to train and validate our models. Then, the round 2 dataset, with 2033 records, will be used to form predictions. Note that “round 1 data” or “training data” refers synonymously to the model training phase, while “round two” or “prediction data” refers to the subsequent application of the model for prediction.

The preprocessing involves feature cleaning and feature engineering steps. As the goal is to mitigate the need for human intervention, many of the sub-steps are designed to be semi-automated by relying on a few generalized assumptions.

4.2.2 Cleaning

Cleaning data is an essential preprocessing step, aim to minimize noise and errors that could be introduced into the training model. Some of the cleaning procedures performed include removing duplicated features, handling missing values, and defining or redefining the data type for certain columns.

4.2.3 Feature Creation

Feature creation involves creating or extracting new features from the dataset. We divide this process into two parts: one aimed at structured data, such as the questionnaire, and the other aimed at text using NLP techniques, such as open narrative, physician keywords, and residual unprocessed data from the questionnaire.

The first set of features to be processed are the physician-coded CoD labels. This set of features will eventually become target labels, hence they will be separated from other features. The VA final CoD coding process is as follows: Two physicians independently submit up to two rounds of their CoD predictions. At any point, if the physicians agreed on the CoD, then that code becomes the final CoD. If the physicians cannot come to an agreement after two rounds, an adjudicator joins in to make the final decision. The CoD codes in the dataset conform to the World Health Organization (WHO) International Classification of Disease Revision 10 (ICD-10). The ICD-10 code is then converted to Centre for Global Health Research for ICD-10 (CGHR-10) titles, which are CoD labels in plain language and will be the final target labels.

First, we need to separate physician-coded CoD labels. As described in the Dataset section, the final CoD is converted into CGHR titles and will serve as the final target labels.

Most features are categorical which needs to be converted into a format usable by ML models. One-hot encoding is used to convert all the values within a selected column into their own respective binary columns. Since this technique creates as many new features as there are unique values within the feature, we only applied this technique to features with ten or fewer unique values. In some cases, features with delimited values and isolated features were also encoded using this method.

Some features, where respondents can provide a wide variety of answers, cannot be efficiently binary encoded but may still contain useful information. Our strategy is to convert them into textual data and treat them using NLP techniques. For each of the remaining features, its values are prefixed with the feature's name to maintain some context. Then, all these features are concatenated into a paragraph to form one feature called "residual". This residual feature is then ready for NLP processing.

The NLP techniques employed targets the open narrative and physician keywords, along with the newly created residual feature. For textual data, we employ two NLP techniques: term-frequency inverse-term-frequency (TF-IDF) and n-grams. TF-IDF

calculates the importance of a word by counting its frequency relative to the entire corpus, producing a numerical value between 0 and 1. It was designed to show the prevalence of a given term within a document and among all other documents. The other NLP technique, *n*-grams, counts the occurrences of words, where *n* is the number of words combined. In this study, *n* is defined to two, creating new features by counting the occurrence of single words (unigrams) and pair-words (bi-grams).

Before applying TF-IDF and bi-gram methods, it is best practice to first “clean” the text to standardize the data by reducing words to their canonical form. his text cleaning process includes converting all words to lowercase, removing stop words, and reducing each word to its dictionary form using lemmatization. The newly created features are then concatenated back with the data that has been preprocessed.

4.2.4 Feature Reduction

Feature creation resulted in hundreds of thousands of new features. Since many parts of the ML pipeline require training and predicting models with an identical number of features and data types, we perform a data reconciliation step. This involves retaining only the intersecting features of the expanded datasets. This is the last step to be applied to both the training and prediction datasets; subsequent steps are only performed on the training dataset unless otherwise noted.

A min-max scaler is first applied to the features of the training dataset to ensure all features contribute equitably. Training and prediction datasets need to utilize similar scaler. Therefore, the fitted scaler from round 1 dataset is saved and applied later to round 2 dataset in the 4.2.7 Model Prediction stage, right before generating predictions.

Since some features might not provide meaningful information to the model and might introduce noise or biases, we have employed some common countermeasures to mitigate its effects.

Features that are non-numerical and have an exceptionally large number of unique values relative to the number of rows might represent date and time, or unique

identifiers and are features that add noise to the model. To address this, we set the threshold to drop such features when the ratio exceeds 0.8. Additionally, one of any two paired features with a correlation greater than 0.95 is also dropped.

Features of all data types with exceptionally low to zero variance suggest that they might not contribute meaningfully to the model. We address this by applying a variance threshold in two passes. In the first pass, all features with a variance threshold below 0.001 are dropped. In the second pass, we drop all features except those derived from TF-IDF with a threshold below 0.01. TF-IDF values are often sparse and may have low variability while still carrying meaningful information. The frequency of terms across the corpus may result in many values close to zero, especially for rare terms. Hence, we employ a less restrictive threshold for TF-IDF derived features.

4.2.5 Feature Selection

At this point, there are approximately 5000 features. Despite reducing the number of features, having too many or too few can still cause the model to overfit or underfit. We opted for the method SelectKBest using the Chi-squared function to rank all the features. The rank scores are sorted in descending order and exported, which will later be used in the hyperparameter tuning stage to identify the most optimal number of features (*K*) to use.

4.2.6 Model Training

Several factors can affect the performance of training a model, namely the number of features used, predefined parameters to optimize, and models employed. The models used were Logistic Regression (LR), Random Forest (RF), Support Vector Machine (SVM), and Gradient Boosting Classifier (GBC). We have also included two automated ML tools, TPOT and AutoSKLearn, in subsequent steps since they are auto-tuned and may serve as a reference for our manually-tuned models.

The training and validation process is as follows: each model is trained repeatedly using K features {k: 10, 25, 50, 100, 200, 500, 1000, 2000, 4000}. For example, LR with 10

features, LR with 25 features, and so on. This process results in multiple k-featured models for each model type. For each k-featured model, a predefined list of parameter values is tested exhaustively using a five-fold cross-validation framework. This involves randomly partitioning the data into five subsets, training on four subsets and validating on one, cycling through all partitions to compute the average performance. The metric used to evaluate performance is accuracy. Each parameter combination produces an average accuracy, and the combination yielding the highest accuracy will be the representative parameter for that k-featured model. Parameters and predicted validation labels will be saved for further analysis as candidates for the final prediction model.

Traditionally, in a cross-validation framework where the dataset is divided into five parts, one part is allocated for validation and testing, while the remaining parts are used for training. However, in our study, we utilized the round 1 dataset twice: initially during the hyperparameter tuning stage to extract the best parameters and subsequently for model training. This approach of using round 1 dataset twice deviates from normal practice. However, there are several rationales that may justify this approach.

A primary goal and challenge in CoD prediction models is ensuring robust performance on future data. The role of the round 1 dataset is to train and prepare the model for future, unseen datasets. While the model is initially trained and potentially biased towards the round 1 dataset, its ultimate test lies in its performance on the round 2 dataset, which is separate from the training and hyperparameter tuning processes. Thus, the integrity of the final evaluation is preserved.

Furthermore, although the dataset originated from the same project, the round 1 and round 2 datasets are temporally separated. Utilizing all round 1 data for training allows us to maximize the amount of historical data the model can learn from, which can potentially improve its ability to generalize to future data. Data collection is an expensive and time-consuming exercise. A common and practical approach in such scenarios is to use all available data up to a certain point of time for training, and then test on future datasets.

In the subsequent analysis of the final predictions, we noticed a slight decrease in PCCC of 0.02 to 0.05 from the validation phase to the prediction phase. It is common to see some performance loss when model is applied to new dataset. Since the loss is small, it suggests the model generalize well and is robust to new data. As more data from the HEAL-SL becomes available, we can continually monitor the model's performance using historical data to optimize the model while maximizing the utility of all available datasets.

4.2.7 Model Prediction

From the previous training and validation steps, each model produced a PCCC value for each set of k-features validated. For each ML model, we identified the hyperparameters corresponded with the highest-performing model. This approach enabled us to build a portfolio of models, which were then retrained on the entire round 1 dataset and subsequently used to make predictions on the round 2 dataset.

Additionally, we incorporated two automated ML libraries: AutoSKLearn [48] and Tree-based Pipeline Optimization Tool (TPOT) [49]. These libraries employ advanced algorithms capable of autonomously determining hyperparameter tuning and selecting the most appropriate models. Similar to the manually methods, we designated the round 1 dataset for training and the round 2 dataset for making predictions using these libraries. This approach provided a benchmark for comparison between manual tuning and automated methods.

The results are label predictions, which are compared with the actual target labels to formulate confusion matrices, classification reports, PCCCs, and other related metrics for evaluation.

5. Results and Evaluation

In our results analysis, we will use both statistical and domain-specific metrics to evaluate model performance. As a recap, PCCC is a domain-specific comparison

metrics that measures the degree of agreement between actual and predicted CoD on an individual-level with added flexibility by allowing minor variations by chance; F1-score is a statistical metric used to measure the harmony between precision and recall; and CSMF accuracy is a domain-specific metric used to measure the agreement between predicted and actual CoD distribution on a population level.

Model	*k*-features	Trained PCCC (rd1)	Predicted PCCC (rd2) ↓	PCCC difference	F1-Score (rd2)	CSMF Accuracy(rd2)
LR	1000	0.6338	0.6153	-0.0185	0.6101	0.8824
SVM	1000	0.6218	0.5989	-0.0228	0.6007	0.8874
A-SKL[§]	2000	N/A	0.5974	N/A	0.5861	0.8401
RF	1000	0.6179	0.5790	-0.0390	0.5621	0.8231
TPOT[§]	4000	N/A	0.5775	N/A	0.5646	0.8236
GB	2000	0.6048	0.5511	-0.0538	0.5413	0.7544

Table 1: Comparative Analysis of Various ML Models and AutoML Techniques. “Trained PCCC (rd1)” represents the average PCCC calculated for all Causes of Death using the round 1 dataset using the round 1 dataset with the best hyperparameters. “Predicted PCCC (rd2)” indicates the average PCCC for all predicted CoD using the round 2 dataset. “PCCC difference” indicates the change in PCCC values, calculated by subtracting the Trained PCCC from the Predicted PCCC. Additionally, F1-Score and Cause-Specific Mortality Fraction (CSMF) Accuracy are presented as performance metrics, calculated based on the predictions for round 2 dataset.

Table 1 shows the performance comparison across ML models in this study. Using the round 1 dataset, all manual models attained a trained PCCC of 0.60 to 0.63, with LR being the highest. Using the round 2 dataset, predictions yielded a PCCC of 0.55 to 0.61, indicating a slight PCCC loss of 0.02 to 0.05. The F1-score is also very close to PCCC values, suggesting that false predictions are not excessive. CSMF accuracy is also fairly high, ranging from 0.75 to 0.88. The performance ranking order for all models is similar for trained PCCC, predicted PCCC, and F1-Score. Interestingly, the loss, calculated as trained PCCC minus predicted PCCC, increases as the model performance decreases. Two automated methods, Auto-SKLearn and TPOT Classifier, which were added as reference comparisons shows performance slightly worse than SVM, but better than GB.

Since LR with 1000 features is our best-performing model, we extend our analysis to exploring the properties of these features. Appendix C shows the data properties of the top 1000 features scored by the Chi-squared function in SelectKBest. In terms of feature source, approximately half originated from the questionnaire of the VA, and a quarter each from physician input and open narrative. Moreover, in terms of feature datatypes, almost three-quarters were features generated by NLP techniques, a little more than a quarter were categorical features by one-hot encoding, and a few were numerical.

We further examined the top 25 features, as depicted in Appendix D, and found that most are related to maternal, injury, and accident-related features except for “diagnosis asthma”. Most of the features originate from the questionnaire with a handful from physician input. Most features are categorical, except a quarter were created using NLP techniques.

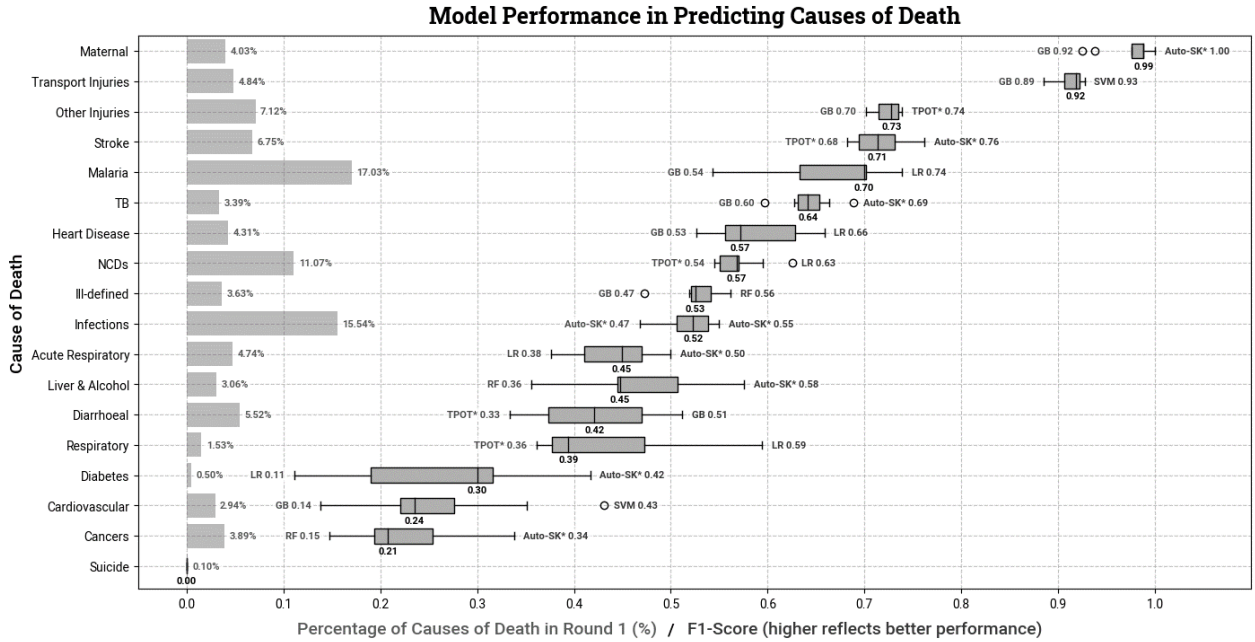


Figure 1 Model performance in predicting each causes using round 2 dataset, overlaid with CoD distribution in round 1 dataset

In *Figure 1*, the boxplot shows the performance for both manual and automated prediction models for each CoD, and the bar graph shows the distribution of the CoD based on round 1 dataset as a reference. “Maternal” and “Transport injuries” CoDs yield the highest median F1-Score greater than 0.90. “Maternal”, “Transport injuries”, “Other injuries”, “Stroke”, “TB”, “NCDs”, and “Ill-defined” have a narrow overall spread of F1-Score less than 0.1, suggesting the models are stable. Whereas “Malaria”, “Liver & Alcohol”, “Respiratory”, “Diabetes”, “Cardiovascular” have a wide overall spread of F1-Score equal or greater than 0.2, suggesting the models are unstable. “Diabetes”, “Cardiovascular”, and “Cancer” have a low median F1-Score equal or less than 0.3, and “Suicide” yielded 0, suggesting models predicted poorly on these causes of death.

When observing the two graphs in tandem to see whether the number of CoDs in the dataset is related to the predictability of the models, we fail to observe a strong association. “Malaria” has the highest percentage of CoD, and although it has a moderate median F1-Score, its overall spread is large at 0.2. “Infection” is the second highest percentage of CoD, with a low overall spread of just 0.08, yet its F1-Score is mediocre at 0.52. The two CoDs with a median F1-Score greater than 0.90 have at least 4% of CoDs in the round 1 dataset, which may suggest a minimum threshold percentage of records needed to create a robust model for a particular cause. “Suicide”, having only 0.1% of cases (five cases to be exact), resulted in a median F1-Score of 0, suggesting insufficient records hinder the model’s ability to predict a particular cause. The second lowest is “Diabetes” with only 0.5% of round 1 data, also resulting in poor model predictability with the largest overall spread of 0.31 and a poor median F1-Score of 0.30.

Of the 18 causes of death, for manually tuned ML methods, LR performed best for four CoDs, RF for one CoD, SVM for two, GB for one, and as for automatically tuned ML methods, Auto-SKLearn for eight CoDs and TPOT for one.

6. Discussion

6.1 Advantages

6.1.1 Explainability

The ML models employed in this study require preprocessing and fine-tuning. While automated ML methods, such as AutoSKLearn, demonstrated better performance for certain CoDs and do not require such extensive manual intervention, there are some trade-offs to consider. As mentioned earlier, AutoSKLearn was the top performer for most CoDs, with LR often ranking second. Although manually tuned ML models demand additional optimization steps, they generally offer greater transparency by providing clearer visibility into model’s inner workings. In the case of LR, after fitting the training data to the model, the coefficients can be accessed to observe the influence of

each feature. The sign and magnitude of these coefficients provide insights into how the features affect the model's prediction. This information can be indirectly used to support understanding feature importance. On the other hand, automated ML simplifies the optimization process and while achieving remarkable prediction performance, it lacks ease of examination of library's parameters. This is due to their tendency to use a combination of methods and models, making them more complex to interpret. Additionally, AutoSKLearn searches through a broad range of parameters and variables for optimal tuning, which prolongs the overall computation time.

6.1.2 Adaptive Data Processing

ML models are generally more flexible and adaptable compared to tools like InterVA and InSilicoVA, which are well-known in the industry and specifically created for CoD prediction. These VA tools require the input data to adhere to certain formatting requirements, adding an additional layer of time-consuming data cleaning and preparation. In contrast, our case study employed automated preprocessing techniques based on few generalized assumptions, as detailed in the Feature Creation and Feature Reduction sections. While our preprocessing is not completely automated, it significantly reduces the manual effort required to prepare the data to be processed.

Furthermore, for features beyond the capabilities of the automated preprocessing techniques, we converted them into text and processed those features using NLP techniques as a fallback solution. As demonstrated, the models adapted well to features engineered in this manner, with three-quarters of the features used in the best performing models consisting of features derived using NLP techniques. This demonstrates the considerable flexibility of self-trained models in utilizing data, alleviating the constrain of being unable to process a particular record due to missing data, and maximize the utility of all available features in a dataset.

6.1.3 Enhanced Control through Model Ownership

The performance of ML models is generally influenced by the availability of sufficient training data. Collecting VA records is time-consuming, as it involves fieldwork

conducted manually. However, as more iterations are completed over time, the repository of datasets will grow, thereby enhancing the performance of models trained on these datasets. In our study, we used the round one dataset for training and the round two data for predictions. In future iterations, we plan to continuously train on historical rounds of data and predict using the most recent data.

Furthermore, as SKLearn is a general-purpose ML library that supports a wide range of ML models, establishing an ML pipeline through SKLearn enhances the flexibility to integrate additional models as needed. Training and predicting on multiple ML models provides access to various outputs, which potentially allows for the combination of results, such as through a voting or weighting system, to create an ensemble of models for predicting CoD. Given that no single ML model has demonstrated significant superiority over others for all CoDs, this approach may prove to be particularly beneficial.

Therefore, this demonstrates that self-trained ML models promotes model ownership, which may provide better control, faster improvements, and more robust models, as opposed to relying on version releases or updates from VA tools publishers.

6.2 Disadvantages

There are several disadvantages and limitations that were observed in this study.

6.2.1 Large Feature Size

ML models utilize vast amounts of data for training, both in terms of feature size and sample size. The round 1 dataset originally contained approximately 360 features. However, encoding structural data features and transforming textual data using NLP techniques significantly increased this number to over 200,000. Feature reduction and selection subsequently brought this number down to approximately 5,000, followed by further optimization through experimenting with the k number of best features to train the most effective model. However, this process also highlights a significant disadvantage: ML models frequently work with large datasets, and all this data,

combined with complex algorithms, inherently require immense computational resources, especially during hyperparameter tuning and model optimization. During our case study, we observed that some methods and models consume more memory and computation time as the feature size increases. Additionally, running our models in parallel utilized extra memory to store data for each instance, presenting a fundamental challenge in prioritizing the program for speed or space.

Traditionally, we mitigate this issue by controlling the number of features we feed into the model, selecting only the most distinguishable ones. However, with developments in NLP, some new engineering techniques, such as embeddings, represent words in a multi-dimensional space. Each token (typically equating to one word) is transformed into a vector that may consist of hundreds of dimensions. Then, each dimension of this vector is converted into a separate feature when used as input. Similarly, even older NLP techniques, such as N-grams used in this case study, also easily generate a vast number of features depending on the n value. In our case study, we limited N to 2—bi-grams. Given W as the number of unique words in the corpus, bi-grams theoretically creates up to W^2 new features, and trigrams creates up to W^3. Hence, moving from bi-gram to tri-gram potentially creates up to W times more features.

6.2.2 Model Runtime Complexities

Aside from complexities introduced by feature size, the choice of model in terms of resource requirements also need to be considered. In our case study, training various LR models was observed to be much faster than GB models, averaging 50 seconds for LR compared to 223 seconds for GB. One reason is that LR is relatively straightforward, using gradient descent to calculate convergence and find the best coefficients. The simplicity also makes parallelization easier as each process can be independently executed in parallel. In contrast, GB builds an ensemble of decision tress iteratively to correct the errors of previous ones, which is inherently more time-consuming. Parallelization becomes more complex when new trees depend on previously built trees. Therefore, effectively using ML for CoD prediction requires a certain level of

understanding of the nuances in ML and computational complexity. Although an inadequately optimized ML pipeline may still perform, the underlying inefficiencies can become more pronounced as the dataset grows larger.

7. Limitations

The two datasets used in this study were collected in separate temporal phases. Data collection practices underwent some modifications in round 2, including changes in feature names and the addition of questions related to COVID-19, resulting in datasets that are mostly, but not completely, identical. After expanding the features, we performed a data reconciliation step to ensure that both the training and prediction datasets used the exact same set of features, including the data types of these features. It was observed that certain categorical features in round 1 had more extensive values than in round 2. As most categorical features are encoded using one-hot encoding, values present in round 1 but absent in round 2 resulted in missing features. Consequently, the data reconciliation step, which retains only intersecting features, would remove these non-matching features. This approach represents a limitation that will need to be addressed in future iterations of data collection. As more rounds are published, continuing to use this method could lead to a decrease in the utilizable feature set. To mitigate this issue, methods such as data imputation or adjustments to the encoding approach could be considered.

As observed in Appendix B, the “Nutritional” CoD did not exist in the round 1 dataset but appeared in round 2. Furthermore, as indicated in Figure 1, “Cancer” yields an F1-Score of 0. These observations suggest two related limitations of such models: the necessity of having sufficient training data to make predictions on any given CoD. In the case of “Nutritional” as CoD, there was no data available in round 1 to train the model, hence the model was unable to recognize this CoD. For “Suicide”, there were only five records in the training dataset, representing only 0.10% of the entire round 1 dataset, which was evidently insufficient for any ML models to make reliable predictions. In the case of LR, using a one-vs-rest classification means an individual ML model is trained for each

CoD, and the absence of a CoD in the training data means no model is trained for the particular CoD. Further research is needed to effectively handle unseen classes, perhaps by means of negative sampling, which is especially crucial in public health where new disease can emerge.

8. Conclusion

We compiled domain knowledge on existing VA methods and processes and developed a conceptual framework for public health professionals, outlining the path to implementing an automated, predominantly ML-based VA process. This framework covers six components: VA data, preprocessing, modeling, model optimization, evaluation, and application. A case study using this framework was employed using VA data from HEAL-SL to train models with round 1 data and predict the CoD using round 2 data. The trained models achieved a PCCC between 0.60 and 0.63, while the predictions ranged from 0.55 to 0.62 in PCCC. This slight loss from training to prediction, along with the temporal separation of the datasets, demonstrates the models' robustness. Logistic regression models achieved the highest average PCCC, while AutoSKLearn came in second in average PCCC but performed best in most CoD predictions. “Maternal conditions” and “Transport injuries” obtained the highest F1-scores of over 0.9. Given these performance, the models are unable replace the role of physicians in VA coding, but can serve as a second opinion to inform physicians of potential alternative CoDs, which is a functionality that can be integrated into the physician CoD coding software platform.

The most effective models in our study predominantly utilized features derived through NLP techniques, applied to both structured and unstructured data. This highlights the potential of text-based features and suggests a need to reevaluate data collection methodologies to be used with ML models, particularly with NLP. The lower utilization of structured data raises questions about the efficiency of extensive questionnaires, hinting at a potential shift towards less structured data in future research. Additionally, while the NLP techniques employed were relatively basic compared to more advanced methods

like embeddings and large language models, they proved to be effective. Since feature engineering is a fundamental aspect of ML models, further research into extracting high-quality NLP-derived features could be highly beneficial. In one instance, we attempted to extract embeddings from VA textual data using BERT-based models. However, this approach quickly generated an extremely large set of features, exceeding the capacity of our existing resources for model training within a limited timeframe—this presents an avenue for further exploration. Other future research opportunities include determining which model performs better with specific CoDs. The concept is to use the best-suited model for each CoD. However, the challenge lies in identifying the appropriate model when the CoD is unknown.

By demonstrating the effectiveness of incorporating both structured and unstructured data in the VA process, this study lays the groundwork for exploring more sophisticated NLP methods in conjunction with ML models to enhance performance in public health applications. Future explorations in this field have the potential to significantly improve the utility of VA. Although perfect accuracy may not be achievable, the application of these advanced methodologies can still play an advisory role to physicians, potentially lessening their workload and thereby improving the overall quality of CoD assignments. Ultimately, such enhancements contribute to more informed public health policies through the provision of reliable CoD data.

References

[1] D. Chandramohan *et al.*, “Estimating causes of death where there is no medical certification: evolution and state of the art of verbal autopsy,” *Glob. Health Action*, vol. 14, no. Suppl, p. 1982486, 2021, doi: 10.1080/16549716.2021.1982486.

[2] P. Bailo, F. Gibelli, G. Ricci, and A. Sirignano, “Verbal Autopsy as a Tool for Defining Causes of Death in Specific Healthcare Contexts: Study of Applicability through a Traditional Literature Review,” *Int. J. Environ. Res. Public. Health*, vol. 19, no. 18, p. 11749, Sep. 2022, doi: 10.3390/ijerph191811749.

[3] S. O. Danso *et al.*, “Population cause of death estimation using verbal autopsy methods in large-scale field trials of maternal and child health: lessons learned from a 20-year research collaboration in Central Ghana,” *Emerg. Themes Epidemiol.*, vol. 20, no. 1, p. 1, 2023, doi: 10.1186/s12982-023-00120-7.

[4] C. Menéndez *et al.*, “Limitations to current methods to estimate cause of death: a validation study of a verbal autopsy model,” *Gates Open Res.*, vol. 4, p. 55, May 2021, doi: 10.12688/gatesopenres.13132.3.

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

[6] H. D. Kalter, J. Perin, and R. E. Black, “Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death,” *J Glob Health*, vol. 6, no. 1, p. 010601, 2016, doi: 10.7189/jogh.06.010601.

[7] A. D. Flaxman, J. C. Joseph, C. J. L. Murray, I. D. Riley, and A. D. Lopez, “Performance of InSilicoVA for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards,” *BMC Med.*, vol. 16, p. 56, Apr. 2018, doi: 10.1186/s12916-018-1039-1.

[8] B. Gilbert *et al.*, “Multi-Cause Calibration of Verbal Autopsy–Based Cause-Specific Mortality Estimates of Children and Neonates in Mozambique,” *Am J Trop Med Hyg*, vol. 108, no. 5 Suppl, pp. 78–89, 2023, doi: 10.4269/ajtmh.22-0319.

[9] M. Garenne, “Prospects for automated diagnosis of verbal autopsies,” *BMC Med.*, vol. 12, p. 18, Feb. 2014, doi: 10.1186/1741-7015-12-18.

[10] S. Danso, E. Atwell, and O. Johnson, “Linguistic and Statistically Derived Features for Cause of Death Prediction from Verbal Autopsy Text,” Springer, 2013, pp. 47–60. doi: 10.1007/978-3-642-40722-2_5.

[11] 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,” *J. Am. Stat. Assoc.*, vol. 111, no. 515, pp. 1036–1049, 2016, doi: 10.1080/01621459.2016.1152191.

[12] E. K. Nichols *et al.*, “The WHO 2016 verbal autopsy instrument: An international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0,” *PLoS Med.*, vol. 15, no. 1, p. e1002486, Jan. 2018, doi: 10.1371/journal.pmed.1002486.

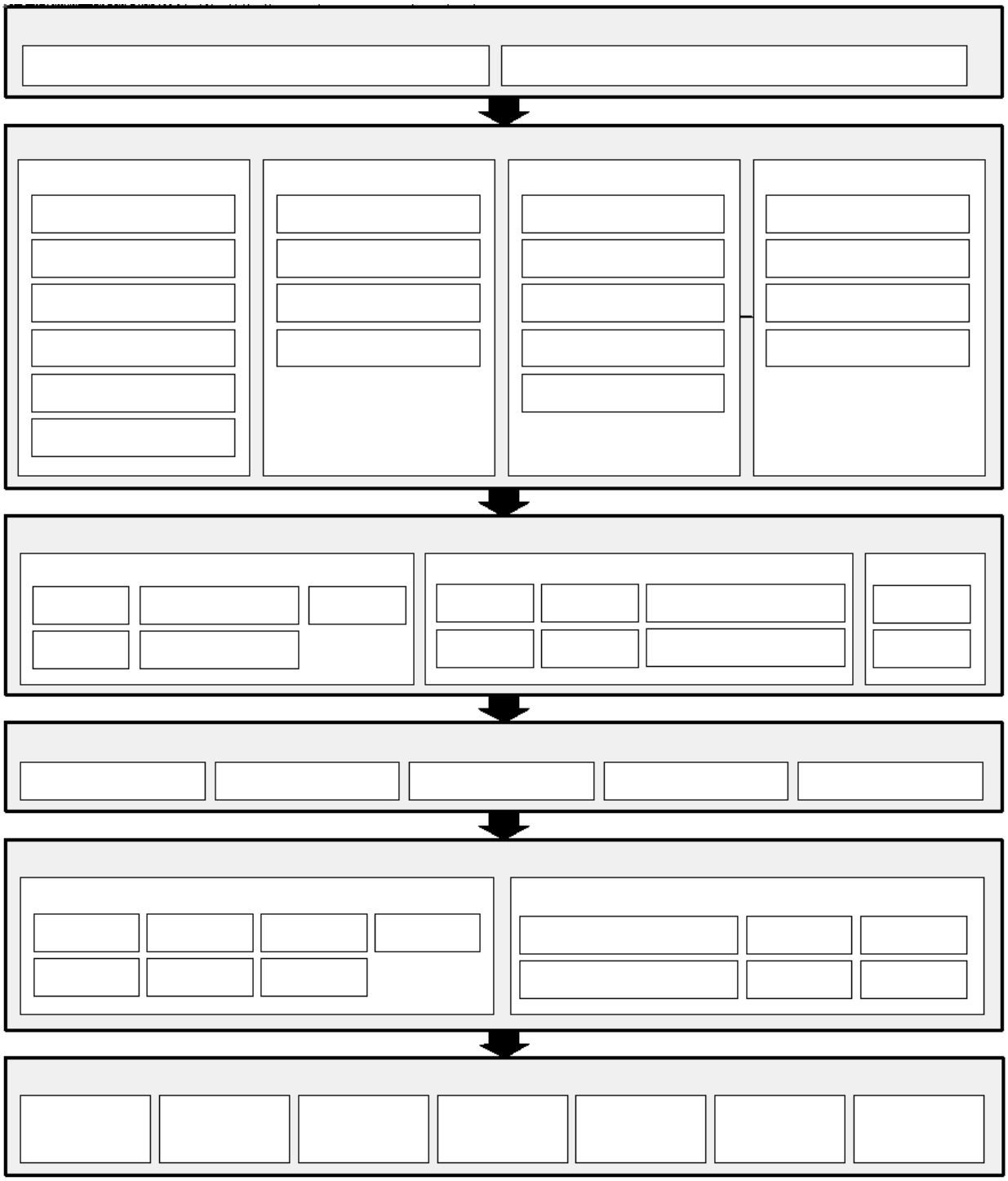
[13] P. Byass *et al.*, “An integrated approach to processing WHO-2016 verbal autopsy data: the InterVA-5 model,” *BMC Med.*, vol. 17, no. 1, p. 102, 2019, doi: 10.1186/s12916-019-1333-6.

[14] M. Tunga, J. Lungo, J. Chambua, and R. Kateule, “Verbal autopsy models in determining causes of death,” *Trop. Med. Int. Health*, vol. 26, no. 12, pp. 1560–1567, 2021, doi: 10.1111/tmi.13678.
[15] G. King and Y. Lu, “Verbal Autopsy Methods with Multiple Causes of Death,” *Stat. Sci.*, vol. 23, no. 1, Feb. 2008, doi: 10.1214/07-STS247.
[16] J. Leitao *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 Med.*, vol. 12, p. 22, Feb. 2014, doi: 10.1186/1741-7015-12-22.
[17] R. Carshon-Marsh *et al.*, “Child, maternal, and adult mortality in Sierra Leone: nationally representative mortality survey 2018–20,” *Lancet Glob. Health*, vol. 10, no. 1, pp. e114–e123, Jan. 2022, doi: 10.1016/S2214-109X(21)00459-9.
[18] M. Hameed and F. Naumann, “Data Preparation: A Survey of Commercial Tools,” *ACM SIGMOD Rec.*, vol. 49, no. 3, pp. 18–29, Dec. 2020, doi: 10.1145/3444831.3444835.
[19] A. Shome, L. Cruz, and A. van Deursen, “Data Smells in Public Datasets,” in *2022 IEEE/ACM 1st International Conference on AI Engineering – Software Engineering for AI (CAIN)*, May 2022, pp. 205–216. doi: 10.1145/3522664.3528621.
[20] S. Velliangiri, S. Alagumuthukrishnan, and S. I. Thankumar joseph, “A Review of Dimensionality Reduction Techniques for Efficient Computation,” *Procedia Comput. Sci.*, vol. 165, pp. 104–111, Jan. 2019, doi: 10.1016/j.procs.2020.01.079.
[21] D. Yogish, T. N. Manjunath, and R. S. Hegadi, “Review on Natural Language Processing Trends and Techniques Using NLTK,” in *Recent Trends in Image Processing and Pattern Recognition*, K. C. Santosh and R. S. Hegadi, Eds., in Communications in Computer and Information Science. Singapore: Springer, 2019, pp. 589–606. doi: 10.1007/978-981-13-9187-3_53.
[22] A. Le Glaz *et al.*, “Machine Learning and Natural Language Processing in Mental Health: Systematic Review,” *J. Med. Internet Res.*, vol. 23, no. 5, p. e15708, May 2021, doi: 10.2196/15708.
[23] S. Danso, E. Atwell, and O. Johnson, “A Comparative Study of Machine Learning Methods for Verbal Autopsy Text Classification,” 2014, doi: 10.48550/arXiv.1402.4380.
[24] S. Idicula-Thomas, U. Gawde, and P. Jha, “Comparison of machine learning algorithms applied to symptoms to determine infectious causes of death in children: national survey of 18,000 verbal autopsies in the Million Death Study in India,” *BMC Public Health*, vol. 21, p. 1787, Oct. 2021, doi: 10.1186/s12889-021-11829-y.
[25] T. Manaka, T. van Zyl, and D. Kar, “Improving Cause-of-Death Classification from Verbal Autopsy Reports.” arXiv, Oct. 31, 2022. Accessed: Aug. 03, 2023. [Online]. Available: http://arxiv.org/abs/2210.17161
[26] A. Datta, J. Fiksel, A. Amouzou, and S. L. Zeger, “Regularized Bayesian transfer learning for population-level etiological distributions,” *Biostat. Oxf. Engl.*, vol. 22, no. 4, pp. 836–857, Feb. 2020, doi: 10.1093/biostatistics/kxaa001.

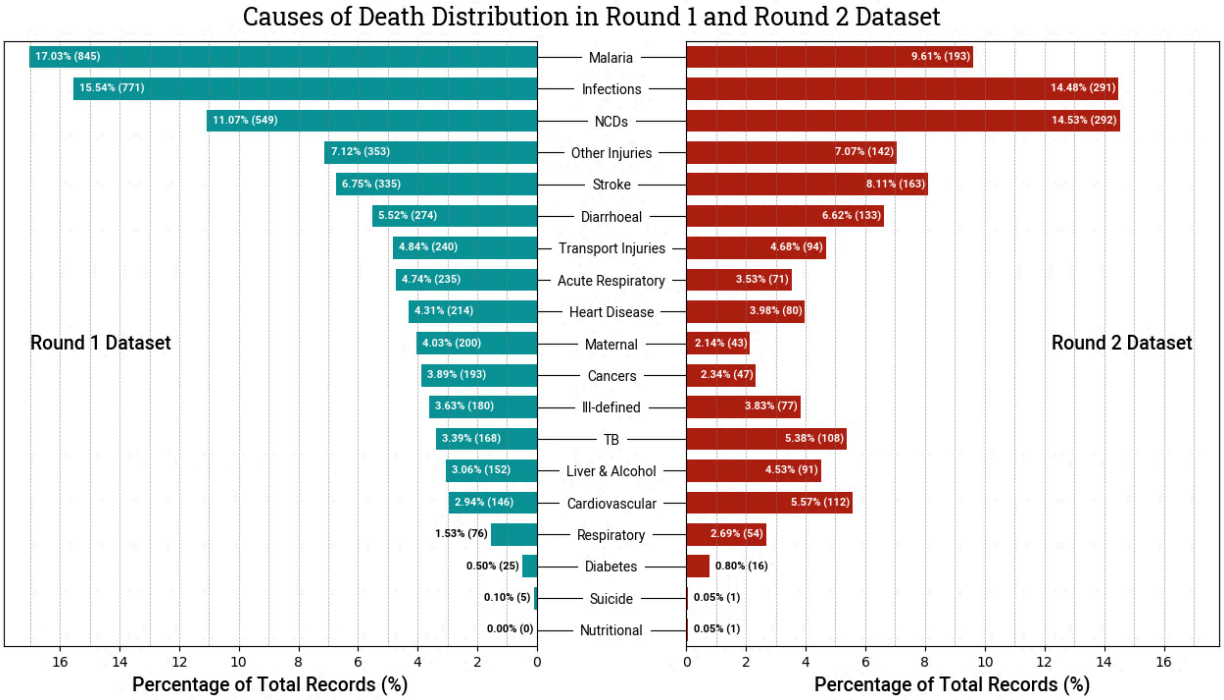
[27] Z. Wu, Z. R. Li, I. Chen, and M. Li, “Tree-informed Bayesian multi-source domain adaptation: cross-population probabilistic cause-of-death assignment using verbal autopsy,” 2021, doi: 10.48550/arXiv.2112.10978.
[28] L.-M. Thomas, L. D’Ambruoso, and D. Balabanova, “Verbal autopsy in health policy and systems: a literature review,” *BMJ Glob. Health*, vol. 3, no. 2, p. e000639, 2018, doi: 10.1136/bmjgh-2017-000639.
[29] P. Serina *et al.*, “Improving performance of the Tariff Method for assigning causes of death to verbal autopsies,” *BMC Med.*, vol. 13, p. 291, Dec. 2015, doi: 10.1186/s12916-015-0527-9.
[30] C. J. Murray, R. Lozano, A. D. Flaxman, A. Vahdatpour, and A. D. Lopez, “Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies,” *Popul. Health Metr.*, vol. 9, p. 28, Aug. 2011, doi: 10.1186/1478-7954-9-28.
[31] R. H. Hazard *et al.*, “Automated verbal autopsy: from research to routine use in civil registration and vital statistics systems,” *BMC Med.*, vol. 18, p. 60, Mar. 2020, doi: 10.1186/s12916-020-01520-1.
[32] D. de Savigny *et al.*, “Integrating community-based verbal autopsy into civil registration and vital statistics (CRVS): system-level considerations,” *Glob. Health Action*, vol. 10, no. 1, p. 1272882, Jan. 2017, doi: 10.1080/16549716.2017.1272882.
[33] C. Ndila *et al.*, “Verbal autopsy as a tool for identifying children dying of sickle cell disease: a validation study conducted in Kilifi district, Kenya,” *BMC Med.*, vol. 12, p. 65, Apr. 2014, doi: 10.1186/1741-7015-12-65.
[34] C. Ndila *et al.*, “Causes of death among persons of all ages within the Kilifi Health and Demographic Surveillance System, Kenya, determined from verbal autopsies interpreted using the InterVA-4 model,” *Glob. Health Action*, vol. 7, p. 10.3402/gha.v7.25593, Oct. 2014, doi: 10.3402/gha.v7.25593.
[35] N. Alam, H. R. Chowdhury, S. C. Das, A. Ashraf, and P. K. Streatfield, “Causes of death in two rural demographic surveillance sites in Bangladesh, 2004–2010: automated coding of verbal autopsies using InterVA-4,” *Glob. Health Action*, vol. 7, p. 10.3402/gha.v7.25511, Oct. 2014, doi: 10.3402/gha.v7.25511.
[36] A. M. Aukes *et al.*, “Causes and circumstances of maternal death: a secondary analysis of the Community-Level Interventions for Pre-eclampsia (CLIP) trials cohort,” *Lancet Glob. Health*, vol. 9, no. 9, pp. e1242–e1251, Jul. 2021, doi: 10.1016/S2214-109X(21)00263-1.
[37] A. R. Saya, J. Katz, S. K. Khatry, J. M. Tielsch, S. C. LeClerq, and L. C. Mullany, “Causes of neonatal mortality using verbal autopsies in rural Southern Nepal, 2010–2017,” *PLOS Glob. Public Health*, vol. 2, no. 9, p. e0001072, Sep. 2022, doi: 10.1371/journal.pgph.0001072.
[38] M. Dheresa *et al.*, “Uncertainties in the path to 2030: Increasing trends of under-five mortality in the aftermath of Millennium Development Goal in Eastern Ethiopia,” *J. Glob. Health*, vol. 12, p. 04010, 2022, doi: 10.7189/jogh.12.04010.
[39] R. Dandona *et al.*, “Distinct mortality patterns at 0–2 days versus the remaining neonatal period: results from population-based assessment in the Indian state of Bihar,” *BMC Med.*, vol. 17, p. 140, Jul. 2019, doi: 10.1186/s12916-019-1372-z.

[40] R. Dandona, G. A. Kumar, A. Kharyal, S. George, M. Akbar, and L. Dandona, “Mortality due to snakebite and other venomous animals in the Indian state of Bihar: Findings from a representative mortality study,” *PLoS ONE*, vol. 13, no. 6, p. e0198900, Jun. 2018, doi: 10.1371/journal.pone.0198900.
[41] R. Dandona *et al.*, “Identification of factors associated with stillbirth in the Indian state of Bihar using verbal autopsy: A population-based study,” *PLoS Med.*, vol. 14, no. 8, p. e1002363, Aug. 2017, doi: 10.1371/journal.pmed.1002363.
[42] J. D. Hart *et al.*, “How advanced is the epidemiological transition in Papua New Guinea? New evidence from verbal autopsy,” *Int. J. Epidemiol.*, vol. 50, no. 6, pp. 2058–2069, May 2021, doi: 10.1093/ije/dyab088.
[43] Md. T. H. Shawon *et al.*, “Routine mortality surveillance to identify the cause of death pattern for out-of-hospital adult (aged 12+ years) deaths in Bangladesh: introduction of automated verbal autopsy,” *BMC Public Health*, vol. 21, p. 491, Mar. 2021, doi: 10.1186/s12889-021-10468-7.
[44] M. Reeve *et al.*, “Generating cause of death information to inform health policy: implementation of an automated verbal autopsy system in the Solomon Islands,” *BMC Public Health*, vol. 21, p. 2080, Nov. 2021, doi: 10.1186/s12889-021-12180-y.
[45] Centre of Global Health Research, “Open Mortality | Help the Living, Study the Dead,” Open Mortality. Accessed: Nov. 28, 2023. [Online]. Available: https://openmortality.org/dataset/heal-sl
[46] “International Classification of Diseases (ICD),” International Classification of Diseases (ICD). Accessed: Nov. 30, 2023. [Online]. Available: https://www.who.int/standards/classifications/classification-of-diseases
[47] P. Jha *et al.*, “Automated versus physician assignment of cause of death for verbal autopsies: randomized trial of 9374 deaths in 117 villages in India,” *BMC Med.*, vol. 17, p. 116, Jun. 2019, doi: 10.1186/s12916-019-1353-2.
[48] “auto-sklearn — AutoSklearn 0.15.0 documentation.” Accessed: Nov. 30, 2023. [Online]. Available: https://automl.github.io/auto-sklearn/master/
[49] “TPOT.” Accessed: Nov. 30, 2023. [Online]. Available: http://epistasislab.github.io/tpot/

Appendix A – Conceptual Framework



Appendix B – Cause of Death Distribution in HEAL-SL Datasets



Distribution of causes of death distribution in HEAL-SL round 1 and round 2 dataset. “Infections” for Unspecified infections, “TB” for Tuberculosis, “Transport” Injuries for Road and transport injuries, “NCDs” for Other noncommunicable diseases, “Cardiovascular” for Other cardiovascular diseases, “Maternal” for Maternal conditions, “Liver” & Alcohol for Liver and alcohol related diseases, “Heart” Disease for Ischemic heart disease, “Diarrhoeal” for Diarrhoeal diseases, “Diabetes” for Diabetes mellitus, “Respiratory” for Chronic respiratory diseases, “Acute” Respiratory for Acute respiratory infections, “Nutritional” for Nutritional deficiencies.

Appendix C – Top 1000 Feature Properties Analysis

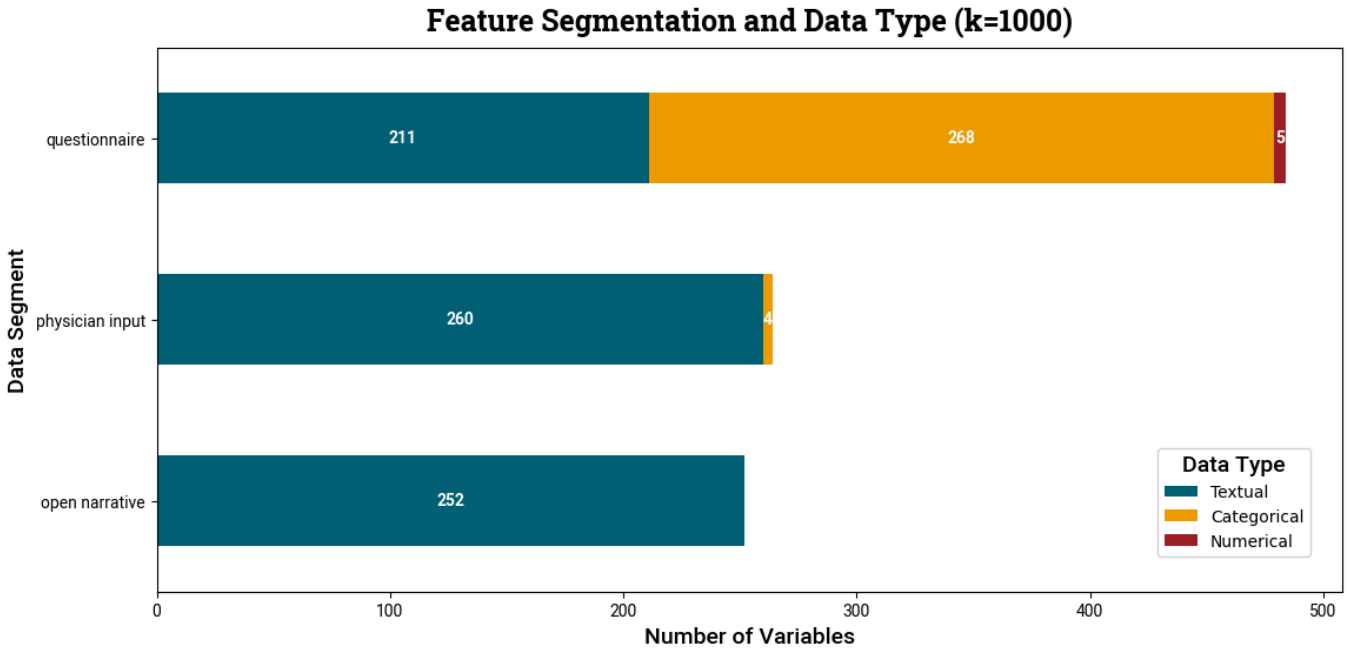
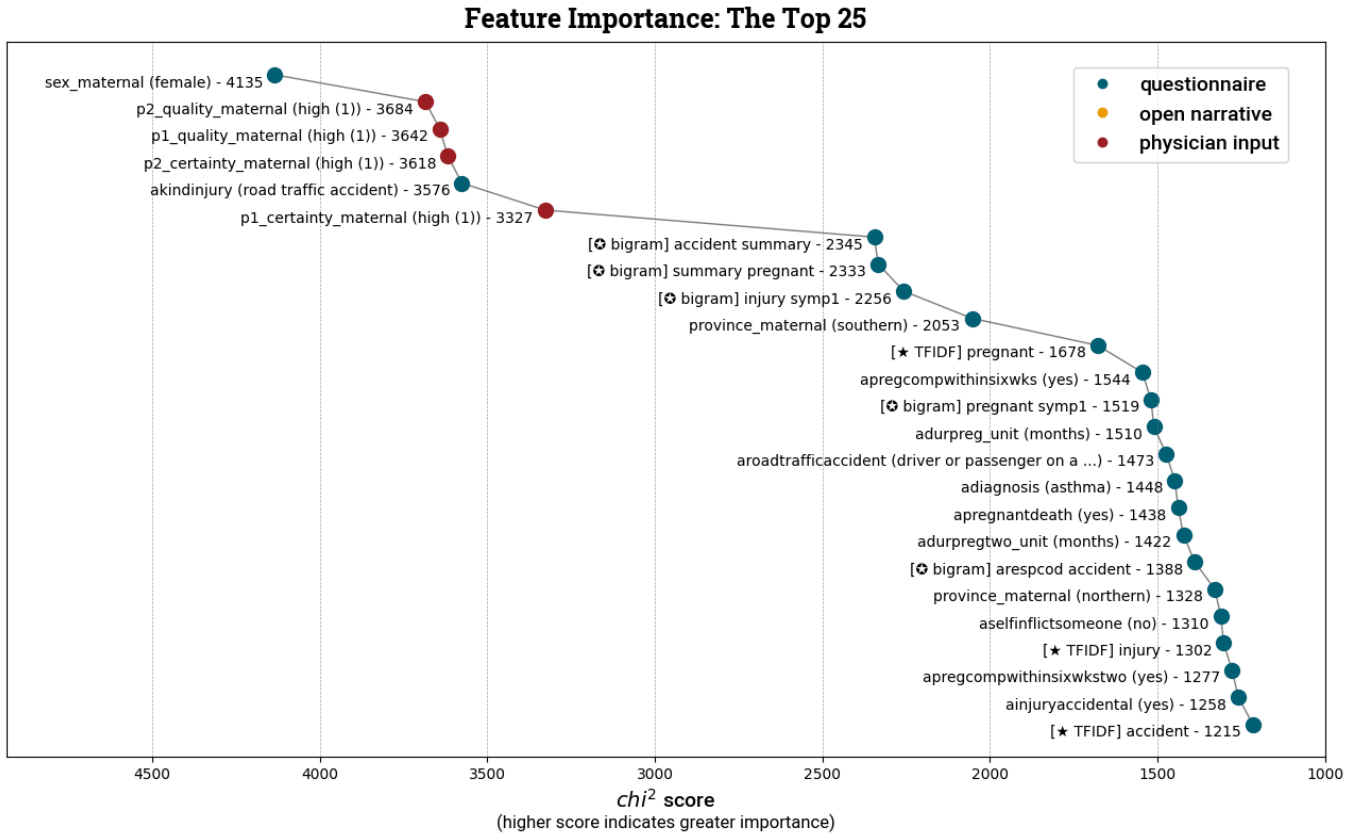


Diagram showing the data type and which segment did the data originated from for the 1000 features

Appendix D – Top 25 Feature Importance Analysis



Feature importance for the top 25 features scored using Chi^2