Overview of Statistical and Machine Learning Techniques for Determining Causes of Death from Verbal Autopsies: A Systematic Literature Review

Michael Tonderai Mapundu (michael.mapundu@wits.ac.za)
University of the Witswatersrand https://orcid.org/0000-0002-2830-0692
Chodziwadziwa Kabudula
University of the Witwatersrand
Eustasius Musenge
University of the Witwatersrand
Turgay Celik
University of the Witwatersrand

Research article

Keywords: Verbal Autopsy, Machine learning, Algorithms, Natural Language Processing, Deep learning, Artificial intelligence

Posted Date: October 26th, 2020

DOI: https://doi.org/10.21203/rs.3.rs-95087/v1

Overview of Statistical and Machine Learning techniques for Determining Causes of Death from Verbal Autopsies: A Systematic Literature Review

Michael Mapundu[a], *Chodziwadziwa Kabudula*[a,b], *Eustasius Musenge*[a], *Turgay Celik*[c,d]

[a] *School of Public Health Faculty of Health Sciences , University of the Witwatersrand South Africa*
[b] *MRC / Wits Rural Public Health and Health Transitions Research Unit (Agincourt)*
[c] *Wits Institute of Data Science , University of the Witwatersrand South Africa*
[d] *School of Computer Science and Applied Mathematics , University of the Witwatersrand South Africa*

Abstract

Background: The process of determining causes of death in areas where there is limited clinical services using verbal autopsies has become a key issue in terms of accuracy on cause of death (prone to errors and subjective), quality of data among many drawbacks. This is mainly because there is no proper standard available in performing verbal autopsy, even though it is important for civil registration systems and strengthening of health priorities. Physician diagnosis is the only gold standard in reviewing verbal autopsy narratives. In practice, conventional statistical methods are used to perform verbal autopsies due to their simplicity and transparency. However, in literature complex machine learning models can be found that can replace the traditional statistical methods. There has not been much application of machine learning techniques in verbal autopsy to determine cause of death, despite the advances in technology. As such, there is a need for a thorough survey of recent literature on statistical and machine learning approaches applied in verbal autopsy to determine cause of death.
Methods: A systematic review was conducted and included a search from six databases. Our study only included scientific articles published in last decade that reported on verbal autopsy and: (1) algorithms; (2) statistical techniques; (3) machine learning and (4) deep learning. The search yielded 110 articles, after meta analysis, we identified 85 articles as being relevant and discarded the other 25. We investigated and compared the most commonly used statistical and machine learning techniques in VAs, identified limitations of each of these techniques, proposed a guiding machine learning framework and pointed to future directions.
Results: Eighty five studies met the inclusion criteria. Apart from physician diagnosis, statistical methods are the most currently applied tools to determine cause of death from verbal autopsies. However, there has been little application of traditional machine learning and emerging techniques, even though they have shown promising results in other domains.
Conclusions: Technological application of machine learning to determine cause of death, should focus on effective ideal strategies of pre-processing, transparency, robust feature engineering techniques and data balancing in order to attain optimal model performance.

Keywords:

Verbal Autopsy, Machine learning, Algorithms, Natural Language Processing, Deep learning, Artificial intelligence.

Corresponding author: michael.mapundu@wits.ac.za

I. Background

Half of the countries of the world do not meet the United Nations 90 percent death registration coverage requirement because deaths in many developing countries 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]. Since cause of death information is required to inform critical health policies and priorities, in the absence of the more traditional, clinically oriented sources, cause of death information should be derived from alternative sources. One widely adopted alternative source of cause of death information is Verbal autopsy (VA). As stated by Thomas et al. [2], the VA concept emanated in Africa and Asia in the 1950's and involves an interview with a relative of the deceased who provides information about signs and symptoms of illness or injury preceding the death and states the cause of death in their opinion. Literature to date suggests that the VA approach was first adopted in India and subsequently research studies using VA surfaced in the late 1970s and early 1980s in Niakhar in Senegal and Matlab in Bangladesh respectively [2]. The use of VA as a source of cause of death information is now common in many developing countries where access to clinical services, death registration and consequently medical certification of causes of death is limited such as Bangladesh, India, Malawi, Niger, South Africa and Nigeria [4].

The process of determining causes of death from VAs is traditionally done through the Physician Certified Verbal Autopsy (PCVA) approach. This approach involves two physicians reviewing the responses to questions from VA interviews to reach a consensus on cause of death, where

they do not agree a third physician decides on cause of death [5]. The PCVA approach is regarded as the gold standard for determining causes of death from VA and is used for training, testing and evaluation of automated approaches for determining causes of death from VA. However, the PCVA has a number of shortcomings as discussed in section IV. In order to improve the determination of causes of death from VA and address some of the weaknesses of the PCVA approach, automated approaches also known as Computer Coded Verbal Autopsies (CCVA) are used as alternative approaches. CCVA methods are data driven and employ automated processing and reasoning. They take as input, the responses to the questions of the VA interviews and maps them to corresponding target ICD-10 cause of death codes or other recognized standard categorization [4; 6; 7]. As argued by Fottrell and Byass [8], one problem with data driven methods of determining causes of death from VA is dependency on the availability of true or medically confirmed causes. Handling of missing data also poses a further challenge as the norm is to treat them as unknown, thus creating a bias in the interpretations.

Driven by recent advancement in technology, there has been a surge in application of computational tools that accurately analyse data for decision making. However, the quality of data is compromised by incompleteness and inaccuracies, amongst other drawbacks [9]. Awareness of these problems is required in research communities in developing and underdeveloped countries [10; 11]. Computational tools use machine learning (ML) techniques that learn from experience relative to some tasks within a class and make use of a performance measure or weighting to improve the performance. Machine learning techniques employ some statistical, probabilistic and optimisation technique in order to discover patterns and trends in complex data through some analysis to get to a decision [12]. There are various statistical and ML techniques that have been adopted in health care to provide required services and address specific problems, such as various classification and regression models that exist to date [13–16]. Jeblee et al. [3] argues that to date, machine learning techniques have been primarily applied to data from only the structured questions of VA interviews, with the best sensitivity scores around 0.60 for individual cause of death classification, using various numbers of cause of death categories. Available evidence indicates that ML can generate real time cause of death information that is similar to that of physicians/experts [17]. Moreover, there has not been full utilisation of deep learning and its capabilities in the health domain, even though they demonstrated great performance and potential in other areas such as computer vision, speech recognition and NLP tasks. Further insights and critique of existing literature on CCVA is discussed in later sections of this review.

In literature, complex machine learning (ML) models can be found that can replace traditional PCVA and some existing CCVA algorithms as approaches for determining cause of death from VAs [18–22]. Even though these ML approaches yield high accuracies in model performance in terms of classification, they lack clear explanations on model interpretability (the inability of the researcher to explain why the machine has predicted the way it has) [17]. There is need to have ML models that are fair, accountable and transparent. Using only hospital based deaths Leitao et al. [5] conducted a study where they compared PCVA and automated methods and concluded that there is no single best performing method out of the two. There is need for further inquiries with computer coded techniques in order to improve performance, thus there is need for publicly available datasets such that comparisons, inferences and deductions can be done.

Subsequently, to the best of our knowledge as of the time of this write-up, there is little evidence that suggest existence of studies that has conducted a systematic literature review on the application of statistical and ML approaches (including deep learning) in VA. Therefore, there is need to have a systematic and comprehensive literature review in order to have an objective understanding of these statistical and ML applications in VA. The review process will highlight various studies in which statistical and ML were applied in VA, thus giving researchers and any other interested parties problems and issues encountered. By so doing this will improve the application of statistical and ML in VA and strengthen health priorities. In spite of high accuracies from ML models; ML models are generally incapable of explaining the predictions due issues of human errors and subjectivity and the issue of fragmented and imbalanced datasets that are taken as input of VA narratives and mapped to corresponding target values on cause of death as per (ICD-10) codes. In this paper, we aim to present a systematic literature survey of statistical and machine learning models which are employed in VA between 2010 and 2020 and propose a guiding ML framework for VA. The remainder of this paper is organized as follows. Section I is the background and has subsections of; feature selection and feature engineering methods, statistical and machine learning models, evaluation metrics, ML optimisation techniques, imbalanced and fragmented datasets, model limitations and assumptions, emerging trends and current technological dynamics in VA and limitations in literature. Section II briefly highlights our methodology. In Section III we present the results and proposed guiding framework for machine learning in VA. Section IV provides a critique of the discussion. Section V is the conclusion. Section VI is a list of abbreviations. All declarations are in section VII. Figure legends are presented in section VIII.

A. *Feature selection and feature engineering*

This section discusses the most commonly used feature selection (FS) and feature engineering (FE) techniques used in VA. The aim of FS and FE is to come up with a vector space that is efficient, scalable and accurate. FS is a process of generating new relevant input features with highest score or weighting from existing ones and FE is creating a new set of features from existing features [23; 24]. FE has three distinct sub-processes namely; feature extraction, feature

value representation and feature selection [25].

*1) **Feature extraction**:* Feature extraction is a process that seeks to extract useful features from a set of reports/narratives or any source of information. Thus feature extraction assigns a document from a predefined set of categories [19]. Various fully automated feature extraction approaches exist in literature [25]. The following extracts discuss various automated feature extraction approaches.

Bag of words (BoW) model extracts unique words from all narrative reports available in the dataset irrespective of their categories and stores in a list (BoW). The BoW technique splits text into words and only includes useful words in the feature vector as dimensions (words) extracted. Each word represents an independent and discriminative feature [25; 26]. However, the BoW ignores the order of words and outputs a generated reduced form of occurrences of each word in text narratives using word frequency [22].

n-grams are a set of co-occurrence words within a given window and are sequential as they make use of continuous number of items such as characters or words from a given sequence of narratives. n-gram models can be of the form; a) n = 1, called unigram, b) when n = 2, it is called bigram, c) when n = 3, it is called as trigram and d) hybrid-grams (combination of unigram, bigram, and trigram) [25].

Word2vec, which can be in either skip-gram and continuous bag of words (CBoW) models. The skip-gram model learns iteratively from existing words available in a sentence to predict the next word. On the contrary, the CBoW model uses the neighboring words to predict the current word. Both approaches employ the parameter window size to determine the limit on number of words[25].

Concept based feature extraction is mainly used to extract medical concepts using medically based ontologies where the relationship between terms is difficult to establish. This is common when there are various terms that mean the same concept. Concept based feature extraction can be done through a bag of concepts or bag of phrases [25].

Graph based or graph of word (GoW) features makes use of content or concept-based features that are represented in graphs to capture word order. In GoW, each given narrative is depicted in a graph structure. The vertex of each graph contains the discriminative feature of the given narrative. Edges are used to connect two co-existing features, where each edge has the weight calculated using the frequency of neighbouring vertices in the graph. The main aim in using GoW features is to consider the word order in the given narratives/ reports [25].

Doc2vec is an approach that applies an unsupervised learning algorithm to come up with vector representations of paragraphs or documents. The algorithm is a modification and application of the concepts of word2vec, and aims at finding similarities in paragraphs or documents [22].

Rule based feature extraction is used in case based reasoning which employs transfer learning (mainly applied when input has few labelled instances that can be used to train an accurate model through transferring capabilities from existing systems to untrained ones). This technique is used to extract features and attributes required by the system's reasoning engine. This can be done through hand coded rules in the information extraction module, that can be searched to attain features and mapped to corresponding values. Conversion can now be done to convert them into a format that the system can utilize for processing. The whole process is automated and makes use of predefined rules [27].

Of all the above discussed feature extraction techniques, Mujtaba et al. [25] suggest the BoW as the commonly used approach for feature extraction. However, the model ignores grammar and word order but word frequency is maintained. To overcome this drawback, n -gram feature extraction technique were proposed. Nevertheless, the n -gram approach does not capture word inversion, ignores the word-level synonymy when applied on clinical text reports and the number of features increases enormously with increasing n, thereby resulting in dimensionality. To overcome these issues, researchers employed other kinds of features such as bag of phrases and bag of contents. These approaches have shown improved accuracy and attained high word level synonymy and polysemy. Although these concept based techniques are useful, nevertheless they still ignore grammar and word order. This has seen the rise of GoW feature extraction techniques which are better in terms of performance to traditional BoW and n-gram models. However, the GoW approaches are computationally expensive [25].

*2) **Feature value Representation Techniques**:* An important step after extracting features from VA narratives is transforming extracted features into numeric vectors, a process known as feature value representation [25]. The aim is to represent extracted features in a master feature vector made up of rows (documents) and columns (features) [19]. Feature representation is done through indexing (allocation of indexing to documents) and weighting (assigning a weight to each term in the document that defines its importance) [23]. There are various weighting schemes and all can take any of the forms below.

Binary representation (BR) defines a feature occurrence/ presence as "1" and non occurrence/not present as "0" [19; 28].

Term frequency (TF) defines the count or number of terms found in a given document. TF gives the weight of a feature based on its frequency in a given document. The more often a term occurs in a document, the more it is representative of the content and its weight should be of high magnitude [26].

Document frequency (DF) is basically the frequency of documents that contain a given term or the count of documents in which that feature occurs [19; 28].

Concept Based Representation is term based frequency that classifies conceptual features as graph based document representation [22].

Term Frequency-Inverse Document Frequency Representation (TF-IDF) is defined as term frequency with inverse class frequency. TF-IDF identifies a term that is a distinctive feature and frequently occurs in a particular

category but is less frequent in other categories, hence has little weight or magnitude [26].

Normalised Term Frequency-Inverse Document Frequency Representation (NTF-IDF) is a technique that uses a normalised factor in relation to TF-IDF as a document length. Terms of same frequency in various documents that apply the normalised factor ensure features derived from short and long documents are of equal importance [28].

*3) **Feature selection**:* Feature selection in VA can be either expert driven or fully automated. The extracts below highlight the various feature selection schemes that exist in VA literature.

Expert driven feature selection also known as PCVA is when a team of experts is specifically responsible for identifying and discovering the useful and discriminative features from the VA narratives. The most frequently counted features are extracted and ranked on the basis of their discriminative power and stored in a vocabulary list for classification [25; 29]. The PCVA process has many drawbacks despite being the gold-standard in VA cause of death determination. The use of scarce various experts to reach a consensus on cause of death, makes the process expensive. Moreover, it lacks consistency as it relies on assessment and review of VA forms [30]. One problem with analysis of VA forms is the variation of performance in the population being studied relative to the true cause of death [30; 31]. Consequently, PCVA is compromised due to poor repeatability of results amongst many drawbacks [4; 32; 33]. PCVA tends to focus on one cause of death even though there might be multiple causes [4]. PCVAs follow own methods when they capture data and they do not have standard questions as features extracted are inconsistent.

Leitao et al. [5] reports on a study where they compared PCVA and CCVA, and they deduced that the Random Forest approach performed better as compared to PCVA methods. Moreover, they conclude that there is no optimum approach for VA coding and suggested further inquiries with large dataset and training of models especially where death is undocumented medically. Mwanyangala et al. [34] investigated completion rates and factors associated with undetermined cause of death and concluded that, there is high completion rate in the initial stages of VA, but a number of deaths are still lost during the later stages of VA process. This is due to physicians not assigning a definite cause of death after receiving the VA forms. Maraba et al. [35] investigated use of PCVA and InterVA and concluded that, the two approaches comparative analysis was slight at individual level and poor at population level, hence they recommended more inquiries and research on VA methods. The wide spectrum of drawbacks above, gave rise to alternative automated feature selection approaches.

*4) **Automated feature selection**:* Automated feature selection entails use of computer programs applying various statistical approaches to automatically extract features from text narratives. These approaches do not need expert intervention. The literature shows that researchers used the automated feature extraction techniques to extract content based features, concept based features, structural features and linguistic features [25]. The next step after feature representation is to create a vector that comprises only selected relevant words useful for categorization, a process known as dimensionality reduction. Various forms of dimensionality reduction are discussed below.

Pearson Correlation (PC), which measures the correlation between two variables thus independent X and dependent variables Y [19]. The correlation coefficient r is between -1 and +1 where -1 denotes a negative correlation and +1 denotes a positive correlation between two instances [25; 29].
Chi-squared test (CS), which measures the statistical independence to determine the dependency of two variables [19; 25]. This statistical test measures the significance level of the observed frequencies of tokens O from the expected frequencies E. Furthermore, this approach is a feature reduction technique where O denotes observed or collected data and E_i refers to expected values.
Information Gain (IG), which measures the importance of a particular attribute in the feature vector space. This metric measures the reduction in uncertainty if the value of the uncertainty is known [19; 29].

Principal component analysis (PCA), which is a statistical method that employs orthogonal transformation to transform a set of observations of correlated features into principal components [25]. In general, the count of principal components is less than or equal to original number of observations. The aim of PCA is to reproduce the correlation matrix using a set of components that are fewer in number and linear combinations of the original set of items. PCA makes it possible to reduce large datasets into smaller ones but avoiding information loss. The benefit of PCA is that it seeks to identify meaningful patterns in a dataset enabling easier exploration and visualization of data [36].
Local Semi-Supervised Feature Selection (LSFS), which is a technique that defines a margin for each data sample in the dataset. It further chooses the most result-oriented features by increasing the margins using a feature weight vector [25].
Fisher Markov Selector (FMS), which is an automated feature selection scheme. It globally selects the optimal subset of features among the classes. This method is useful and effective for handling high dimensional data efficiently [29].
Improved Global Feature Selection (IGFS), which is an ensemble method where the power of global feature selection method and a one-sided local feature selection are combined in a different manner [29].

B. Semi Supervised learning

VA is a semi supervised learning problem. Specifically it is a multi classification problem, where the aim is to classify VA narratives to various disease categories as per ICD-10

codes. In this section we first define semi supervised learning problem and then present most commonly used statistical and machine learning methods in VA.

*1) **Semi supervised learning problem**:* Semi supervised ML approaches seek to optimise classification accuracy with few labelled data points. This approach is relevant in cases where there is a large volume of unlabelled data. However, process of labelling is costly, difficult and labour intensive [25]. On applying semi supervised machine learning approach, the aim is to maximise the training data set that contain the input and output variables x and y, respectively. Thereafter, this training data is provided as an input to the learning algorithm to learn the mapping function $[y = f(x)]$. The main goal of semi supervised learning is to efficiently approximate this mapping function, thereby enabling the accurate prediction of the output variable for new input data x [25].

Classification in VA

A classification method is defined as a description of a specific analytical technique that employs a classifier with predefined rules derived from empirical data applying any classification approach [37]. Classes are derived from classification methods and have attributes that describe them. Classification approaches can fall in any of the following: i) Linear and discriminant techniques which use least squares, hyperplane, quadratic and logistic regression function to separate the dimensional attribute space by fitting a line that maximises categories of data points for each class., ii) Probability density function which applies the k-nearest neighbour approach which is based on the principle that a new case is most likely to be located near-to other cases of the same category in the feature space. Therefore, new cases are classified to clusters according to their class probabilities of their locations in the attribute space and iii) Decision trees and rule based methods which iteratively divides the feature space into smaller regions until branches are pure, thus there will only be branches belonging to a specific single class [37].

Classification requires the creation of a classifier from training data to predict the VA classification/categories [38]. In simple terms in the VA domain, classes can be thought of as validated cause of death and attributes are symptoms/signs that are collected through interviews with the relatives of the deceased. The feature space is made up of observations that have certain conditional probabilities with corresponding scores [37]. Therefore the aim of classification methods is to divide the attribute space into predominant clusters of a single class of observation [37; 39].

*2) **Statistical models**:* Various statistical learning models have been applied in VA domain to date. The following extracts discuss statistical approaches.

Logistic regression

The study by Kocheturov et al. [14] elaborate on logistic regression models as being widely used in biomedicine due to their user friendliness, high level of interpretability and their optimum use in solving problems. In this statistical model we assume $w \in \{0, 1\}$, where 0 represents a bad class and 1 a good class. Therefore, the logistic regression model defines X as the data set and θ as a vector of coefficients derived from maximum likelihood estimation [40].

Naive Bayes

The Naive Bayes classifier (NBC) is based on the Bayes rule and makes the assumption that features in the dataset are mutually exclusive or independent from each other [24]. This implies that the probability each feature contributes, is independent to the final class probability. This approach can be trained efficiently as it assumes value independence of a specific feature from the value of other features, when given a class variable [41]. This statistical model analyses the relationship between each feature of a class and the class to derive a conditional probability between feature values and the class. This computes the probability of instance x given class y assuming the features are independent [21].

Murtaza et al. [42] argue that the NBC has shown better results in terms of VA classification. Miasnikof et al. [21] implemented the NBC in VA as; each VA record or narrative applies NBC to assign a probability of cause of death to each label corresponding to independent conditional probabilities of that label given specific symptoms and the unconditional probability of that same label. Therefore, the label with the highest probability is then assigned as the appropriate cause of death to the VA record. This implies that, we assign each record the cause of death C_{j*} with the highest probability, given a set of n recorded symptoms/ features in the verbal autopsy, denoted by $F1, ..., Fn$.

*3) **Machine learning techniques**:* ML approaches specifically can be powerful, effective and efficient tools in developing better semantic structures to solve issues around automated VA classification [43]. This is evidenced through their superior capabilities such as; use independent parameters which are non parametric, fast in testing, produce results in complex domains, reduce problem complexity, easy to implement and employ more resembling characteristics of the model [24]. Boulle et al. [32] points out that machine learning models can understand better the input-output relationship. Moreover, ML techniques are extensive, comprehensive, highly reliable and consistent relative to changes [4]. The study of Crown [44] concurs with the above assertions and points out that, ML approaches have been employed due to their ability to handle high dimensional data and producing good results with limited training samples.

Boulle et al. [32] applied Artificial Neural Networks (ANN) to classify cause of death from VA narratives. However, they reported issues of over fitting and suggested increasing the number of training examples could improve ANN. Jeblee et al. [3] did a study on VA classification and they reported that there is no technique of automating VA classification that can be better than the PCVA. Albeit, their study only used conventional ML approaches. Pestian et al. [45] applied Natural Language Processing (NLP) for classification as a deep learning model (advanced ML approach). Nevertheless,

they used a small sample size and their results cannot be generalised across studies. Jeblee et al. [46] did a study on multi-task learning for interpretable cause-of-death classification using key phrase prediction through clustering. In spite of that, they reported a low score in terms of accuracy. Rajput et al. [41] used unstructured clinical datasets applying text mining and ML for automatic identification of protected health information and got promising results. Yan et al. [20] did an investigation on application of character embeddings (representing each character in a narrative as a numeric in the vector space) in VA and they reported a low score in terms of accuracy. Regardless, they concluded that application of character embeddings can be useful in VA.

Machine learning approaches have been used effectively for epidemic forecasting including surveillance for health systems. Nonetheless, there is need to focus on real-time monitoring and forecasting [47]. Literature reports on application of conventional shallow ML models which are trained on very high dimensional and sparse features in VA domains [6; 18; 18; 19; 21–24; 26; 28; 29; 31; 45; 46; 48]. We discuss various ML approaches in the following extracts.

k-nearest neighbour

The k- nearest neighbour model derives it's concept from the distance or similarity function for given pairs of observations [24]. The k-nearest neighbour determines similarity function by calculating the Euclidean distance or the cosine similarity measures. However, the best choice of k depends on data. Thus, a good choice can reduce the noise effect [24]. Furthermore, it uses all available instances to categorise a new instance based on the distance function [19]. K-nearest neighbour can also learn by analogy through n dimensional numeric attributes [49]. Therefore, data points which are close to each other are called nearest neighbours. The majority class will be the resultant class of that data point.

k-means clustering

k-means clustering is an unsupervised learning approach for grouping similar objects with similar characteristics together [50]. It is achieved by firstly defining the wanted number of clusters k for n data points, randomly assign each data point to a cluster, compute cluster centroids, re-assign each point to the closest cluster centroid, re-compute cluster centroids and repeat last two steps until no improvements are possible to reach global optima [36]. There are a few studies that have applied this approach to date. One study by Jeblee et al. [46] used multi-task learning for interpretable cause of death classification using key phrase prediction. They however concluded that this approach can improve cause of death classification and can be interpreted easily even though in future they have to improve on accuracy.

Decision trees

According to Mujtaba et al. [19], decision trees 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 point can be generated and assigned a class based on the majority vote of the leaf nodes.

Random Forests

Random Forests has been used in the VA domain because of easier implementation. This approach uses many individual learner trees to perform classification based on votes on an overall category from a given set of inputs [51]. A majority vote is applied to determine a class of a new data point through combining random forests and decision trees. Combined decision trees k are built on randomised algorithm from the bootstrap samples with n observations [31].

Support Vector Machine

The Support Vector Machine (SVM) approach has been applied in the VA field because of the ability to handle high dimensional data, non-linearity and producing good results with limited training samples. This technique employs the idea of a decision boundary known as a hyperplane that distinguishes between classes in a high dimensional space [52]. The documents that are close to the hyperplane are called the support vectors [24]. The SVM approach aims at maximising the margin. SVM computes the maximum linear distance between separated classes in feature space. In non-linear cases the SVM has the ability to use kernels which can map non-linearity between classes, categories and feature space.

Artificial Neural Networks

Artificial Neural Networks (ANNs) which form the basis of deep learning has been applied in VA domain, and uses a neural classifier that can be thought of being a network of units where each unit can be a term and the output unit represents a category [24]. The ANN is made up of the input, hidden and output layers. 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 [24]. 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 [13; 53]. The ANN applies an error function at the output layer to generate updated weights using back propagation (process of fine tuning weights based on error rate computed from previous iteration). Effective and efficient back propagation makes models genaralizable and reliable. ANNs have been applied in VA and have shown promising results even though there is need to have more training examples for the model to be effective [32]. Jeblee et al. [3] applied the various ML approaches and got best results from the ANN using a feed-forward network (no back propagation) with one hidden layer and achieved sensitivity of 0.695. Figure 1 below shows the ANN model of a single neuron in a feed-forward ANN. Input information x_i is received from other neurons, and takes the product of input information and each data point corresponding weight denoted by w_{ij}, and produces a weighted output applying an activation function $f(x_j)$. The weighted output is again passed as input to another neuron

in the next layer, and the same process is repeated until the output layer is reached.

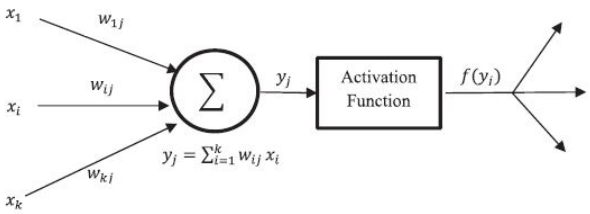

Fig. 1. ANN structure of single neuron.

Decision rule

Decision rule is a classification approach that uses an inference rule based engine to classify documents to their respective categories. Furthermore this approach can entail building a knowledge base with a set of terms or keywords derived from consultations with domain experts [54].

Case based reasoning Case based reasoning (CBR) has been applied in literature by Yeow et al. [27] and employs reasoning through a rule based approach to construct meaning through synthesis of information to derive a new case in an autopsy report. This new meaning constructed through extraction, is transformed into a set of features and values, which are required by the system to perform a similarity analysis. The new generated feature values is compared with the past cases by computing the similarity of each individual feature. It employs the Naive Bayes learner to assign weights to features based on significance and is used as a computation to the final similarity analysis. Similarity analysis is performed using the k-nearest neighbour approach, by computing the similarity scores for all the past cases in the case base. Only top scores will be selected using similarity values and identified by user as the desired outcome, and thus stored in the case base [27].

4) ***Ensemble classifiers****:* The following extract discusses ensemble approaches, which are hybrid schemes of various ML approaches and use a voting scheme or weighting in order to do the classification. These approaches mainly use several prediction tools and aggregate their prediction, such as such as voting, associative classifiers, centroid based classifiers discussed in the study of Korde and Mahender [24]. Implementing ensemble methods improves prediction quality [6; 42]. However, these approaches suffer from very low interpretability and high computational time [14].

Boosting

Boosting also known as an iterative ensemble classifier is a ML approach that improves a weak algorithm into a stronger one through aggregation of multiple classifiers [55]. It employs a weighted vote when labelling unlabelled instances that ultimately becomes a composite classifier [39]. The most common boosting algorithm is The ADABOOST which is short for Adaptive Boosting. The approach has been used in many studies such as [6; 55].

Bagging

Bagging is an approach that builds a classifier for every bootstrap based on n bootstraps from given datasets. Bagging is also known as bootstrap aggregation [39]. Given each trial denoted by $(t = 1, 2, 3...T)$, where the dataset is of size N. The learning environment generates a classifier C^t from the dataset and a final classifier C^* is created from aggregation of the T classifiers that are classified from instances x through voting for class k where it is recorded for every classifier $C^x = k$ and C^* which denotes the class with the most votes.

5) ***VA algorithms****:* VA algorithms have been the alternative traditional method of assigning cause of death apart from the PCVA technique. VA algorithms require VA data derived from real deaths and Symptom Cause Information (SCI) (which is a repository of information about symptoms that are related to each probable cause of death). The SCI may be derived from the physician's reports and logic (that entails a logical algorithm) that combines the two to identify cause-specific mortality fractions (CSMF) to assign specific causes of death. The literature to date identifies five various free online VA computational algorithms that have been widely employed and most of them apply statistical approaches [56].

The InterVA algorithm uses the Bayes rule (formula that computes conditional probabilities of X when given evidence) and Symptom Cause Information (SCI) as conditional probabilities of symptoms given a specific cause of death using Cause Specific Mortality Fractions (CSMFs). These conditional probabilities are sourced from physicians and the Population Health Metrics Research Consortium (PHMRC) gold standard dataset [57]. The PHMRC is responsible for supporting survey design, primary data collection, analysis, development, and to provide tools for quantifying population health. VA datasets cannot distinguish between the confirmed absence of a symptom and missing data. Therefore, to address this limitation of the data, InterVA-4 only utilizes the confirmed presence of a symptom. Moreover, this approach uses propensities (which are specific to a particular cause of death) and not probabilities. Miasnikof et al. [21] applied the Inter VA and reported a sensitivity of 0.43 and CSMF accuracy of 0.71.

PHMRC 'gold-standard' data have both VA and medically certified cause of death. This contains information about the relationship between VA symptoms and medically certified causes. 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 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 [56]. The standardized Tariff Score Q_{ki} for death k and cause i is the quantile value of S_{ki} in the distribution of S_{*i} formed by a set of deaths sampled with replacement from the PHMRC data set so that they have a uniform cause distribution. Values for Q_{ki} are bounded between 0 and 1 so they have the same scale and can be compared. The cause chosen for each death

k is the one with the largest Q_{ki}. James et al. [58] applied the Tariff approach and reported 0.505 Chance-corrected concordance (CCC) (is a measure of how well the predicted cause of death categories correspond to the correct cause of death categories) and CSFM accuracy of 0.770.

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 [59]. This implies that the algorithm uses SCI and derives this based on the Naive Bayes model based on conditional probabilities (symptoms given cause of death and cause of death given symptoms). The probabilities can be sourced from physicians or use a gold standard dataset that can be a de-biased physician-coded VAs. The study by McCormick et al. [7] applied the Insilico approach and reported a mean sensitivity of 0.341 across 34 cause of death categories and 0.85 CSMF accuracy.

The Naïve Bayes Classifier (NBC) algorithm applies the bayes rule to categorise cause of death. The study by Miasnikof et al. [21] applied the NBC and achieved better results as compared to InterVA and the Tariff approach with a sensitivity of 0.57 and CSMF accuracy of 0.88. Nevertheless, their model used only 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 [56]. Desai et al. [60] used the King-Lu and reported a CSMF accuracy of 0.96

Nichols et al. [1] argue that these VA algorithms 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 [56; 60]. Various VA algorithms are dependent on sample size, age group, causes of death, data set size or characteristics of the sample in order to produce best results [4; 61]. 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 [21]. Moreover, the validity of the VA approaches performance in terms of sensitivity, specificity and predictive values vary with regard to causes of death across populations [62]. 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 [21; 56]. 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. Soleman et al. [4] recommended incorporating fully trained multiple translators. A standardised and uniformly reliable method is required to determine cause of death and also a standardised questionnaire which can be applied globally [4].

C. Deep learning

Deep learning approaches have been implemented successfully in literature for the past decades, with the aim of of improving classification. Moreover, deep learning models with optimal hidden layers have been developed to reveal information not easily detectable with traditional statistical and machine learning models. Additionally, deep learning achieves learning of data representation with varying levels of data abstraction when even computational techniques use few processing layers [25]. Therefore, these advanced approaches can be relevant in VA data with high-dimensional sparse data.

Literature reports a wide gap, especially in applying advanced machine learning techniques known as deep learning models as approaches to solving VA problems. Yan et al. [20]; Mujtaba et al. [25]; Jeblee et al. [46] are some of few studies to have used deep learning approaches to date. Yan et al. [20] investigated the application of character embeddings to improve cause of death classification using VA narratives. In their study they applied distributed word vectors (Sentence2vec, GLoVE and Word2vec) and combined two character embeddings techniques (CNN and GRN) for classification and they concluded an improvement in cause of death classification. Application of Natural Language Processing (NLP) as a deep learning approach to solve VA problems has shown promising results [3; 28; 63–65]. NLP is an approach that entails converting human language into formal easy to understand language for computers, and uses a range of computational techniques for automatic analysis and representation of the human language [66; 67]. NLP has efficient, superior performance, integrated feature learning and effective capabilities of attaining end to end learning from complex and multi modality data [43]. NLP word embeddings (representation of text in numeric form stored as vectors) on dense vector representations, enabling multi-level automated feature representation learning, has to date produced excellent results [66]. Friedman et al. [68] used a method based on NLP to extract relevant clinical information and discovered that the approach was better than PCVA diagnosis. Gajalakshmi and Peto [69] investigated use of NLP on suicide note classification and discovered that NLP can assist in the classification process. Kamath et al. [70] did a comparative study of existing ML and deep learning approaches and found out that deep learning was better in terms of performance as compared to machine learning. Recent deep learning methods such as Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short Term Memory (LSTM) and the Transformer architecture have managed to explicitly solve most NLP tasks [43]. The following extracts highlight

the deep learning models that are used in VA. Figure 2 below compares the conventional ANN with one hidden layer and the deep learning architectures with multiple hidden layers.

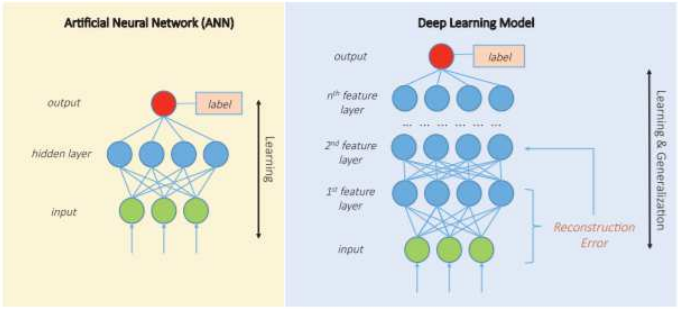

Fig. 2. Comparison of ANN and deep learning architecture.

1) ***Distributed vector representations****:* Distributed representations emanated from the issue of complex and high data dimensionality for traditional statistical NLP. Their end result is the distributed representation of words in low dimensional space. Distributional vectors mark the beginning of data processing layer in a deep learning model and imply that words with similar meanings tend to occur in similar context, thus similarity is measured using the cosine similarity coefficient [66]. The study of Jeblee et al. [46] successfully applied distributed vector representations.

Word2vec

Word2vec is an approach that applies the concept that words within a vector space are of fixed length [71]. Word2vec builds vocabulary from a training text corpus and creates a vector space for each word through learning, and clusters similar words based on their distances using the cosine distance Ma and Zhang [72]. It is made up of two models to develop and construct distributed vector presentations which are; Continuous Bag-of -Words (CBOW) and Skip-gram models. CBOW predicts a word based on its context whilst the Skip-gram model predicts context based on the neighbouring words, thus it discovers semantic relations amongst terms in the corpus [73; 74]. The Word2vec uses the input, projection and output layers [73]. The input layer takes one hot word vector of V neurons and corresponding to hidden layer N neurons. The softmax is output that contains all words in the vocabulary. Each word in the vocabulary is denoted by two learned vectors v_c and v_w where the connected layers use a weight matrix [66]

Sentence2vec

Alami et al. [75] discusses the Sentence2vec which applies the same concept of Word2vec. They argue that a sentence can be seen as text unit and the summary of the document being made up of sentences. They point out that each word in the corpus has an *id* that denotes its position in the dictionary. Sentence2vec represents each sentence in a document as a vector, unlike Word2vec which represents each word as a vector [75].

Global vectors The Glove technique is count based and makes the supposition that the meaning of a word is implied by it's own contextualisation [73; 75]. It is mainly based on word occurrences in a textual knowledge base. This implies that words with similar contexts should also have similar embeddings.

Doc2vec

This technique uses an unsupervised learning algorithm to come up with vector representations of paragraphs or documents. The algorithm is a modification and application of the concepts of Word2vec and it aims at finding similarities in paragraphs or documents [22]. The following extracts present deep learning architectures that focus on character embeddings.

2) ***Character embeddings****:* Character embeddings assume each word to be composed of not more that one character. They can handle the issue of out of vocabulary words (words that appear a few times in the test set but are not in the training corpus of words), unlike distributed vector representations [66].

Convolutional Neural Networks

Convolutional Neural Networks (CNN) is an example of a character embedding technique. CNN was developed as a feature extraction function to extract high level features from n-grams, thus the features would be used for various NLP tasks [66; 76–78]. The CNN is made up of an input layer, lookup table, convolutional layers, pooling layers, fully connected layers and softmax function for classification. After taking input, a look-up table is used for transformation of each word into a vector of user-defined dimensions (e.g. input sequence of $(s_1, s_2, ...s_n)$ of n words is transformed into a sequence of vectors $(w_0, w_1, ...w_n)$. Learning in this network process is done using weights. Data representations and extractions are done on the convolutional and pooling layers. The convolutional layer makes use of a various numbers of filters (kernels), where one kernel extracts a specific pattern of n-gram. The pooling layer employs a maximum pooling strategy which performs sub-sampling on each input typically by applying a maximum operation on each filter. Maximum pooling therefore, maps the input to a fixed dimension of outputs regardless of the size of filters and thus a reduction in the output's dimensionality. Lastly, a fully connected layer employs a feed forward ANN. The fully connected layer consists of an input layer, hidden layer/s and an output layer. At the output layer, a loss/cost function is applied in this context known as Softmax classification [66]. The study of Jeblee et al. [46] successfully applied the CNN and Gated Recurrent Neural network discussed below. [25] argues that the CNN has the ability to learn complex features from the clinical data set compared with conventional statistical and machine learning approaches. Figure 3 shows the visual representations of the CNN architecture.

Gated Recurrent Neural network

Recurrent Neural Networks (RNNs) is a form of character embedding architecture. RNNs are employed in NLP to process sequential information. RNNs produces output that is dependent on previous tasks and results by performing same task on each instance. They make use of memory

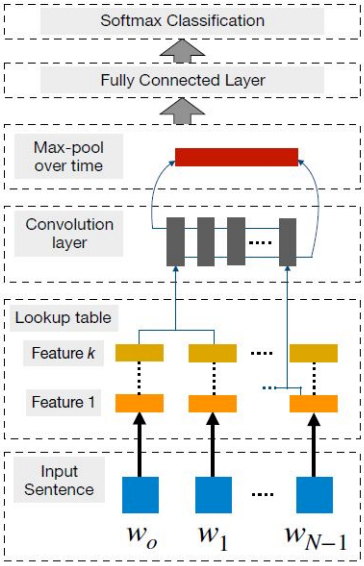

Fig. 3. Convolutional neural network framework.

to pull previous computations for processing with current operations [66]. However, RNNs have issues of "vanishing gradient" (where weighted activation function inputs of a neural network increase or decrease whilst the derivative function approaches zero) which ultimately affects accuracy and makes it difficult for the architecture to learn and hyper tune parameters for optimum results. The Gated Recurrent Neural network (GRNN) has the rest and update gate that control flow of information. This approach further has no control of the hidden layer and has no memory unit. The GRNN has been applied in the study of Yan et al. [20]; Jeblee et al. [46] and showed promising results.

D. ML Performance Metrics in VA

This section will give an overview of the ML evaluation metrics that have been applied in VA to date. More approaches to evaluation, accuracy and validation of algorithmic performance have already been reported on. Nonetheless, in medical data focus is given to Recall than Accuracy [9]. Clark et al. [57] points out that performance evaluation to determine the predictive accuracy of VA algoriths can be done at individual level or population level. At individual level we seek to determine specific cause of death categorisation, whilst at population level we seek to estimate Cause Specific Mortality Fractions (CSMFs). These processes are key to achieve comparability. Assuming a supervised learning problem with a labelled deaths dataset, there is a training set $T = (x1, y1...; x_n...; x_n)$ where each x denotes a symptom or indicator and each y denotes a known cause of death. As such, we have a model of a VA algorithm f to attain the predicted label for a new dataset (X, Y) is denoted by function f. Given the above proposition we can therefore attain the performance of the algorithm that depends on the training data.

A confusion matrix is the benchmark of our evaluation matrices, and consists of True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN) and is used for calculating the metrics which are discussed in this section. Below we discuss various performance evaluation measures in VA.

Actual labels	Predicted labels		
		Positives	Negatives
	Positives	TP	FP
	Negatives	FN	TN

Fig. 4. A confusion matrix with correct predictions, True Positives (TP) and True Negatives (TN), and incorrect predictions, False Positives (FP) and False Negatives (FN).

Macro precision is also known as Precision/ Positive Predictive Value (PPV) refers to the ratio of correctly predicted positive VA reports to the total positively predicted VA reports. This denotes how good a classifier identifies positive observations without much on the negative ones.

Macro recall is also known as Sensitivity or True Positive Rate (TPR) defines the ratio of correctly predicted positive VA reports to the all VA reports in actual positive class. This implies the rate of correctly predicted positive observations or diagnoses (cause of death) [29].

F-measure is a test score to measure accuracy of a particular model. It is a weighted combination of Precision and Recall scores and is computed as the average or mean of the Precision and Recall [29].

Overall accuracy denotes all classes with classified results that have been predicted correctly in fraction terms, thus it is the ratio of correctly predicted VA to the total VA reports [29].

Area under curve (AUC) is employed to evaluate the performance of machine learning algorithms. It plots the rate of true positive (TPR) against the rate of false positive (FPR). The curve evaluates the classifier using 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 is closer to upper left corner and the poor performance of an algorithm is given a weight of 0.5 and below [9; 29].

Specificity is a function that computes the proportion of negative clinical reports that are correctly predicted as a negative [25].

Overall chance-corrected concordance (CCC) defines CCC for cause j as the overall weighted sum of cause-specific CCC [57]. Murray et al. [79] points out that given a proposition of measuring agreement on a cause of death, sensitivity and concordance never considers the agreement expected by chance alone. As such, given a VA algorithm that assigns randomly cause of deaths by chance, we would expect concordance of $\frac{1}{n}$.

Cause Specific Mortality Fractions (CSMF) accuracy defines accuracy as having a value between 0 and 1 and it assumes the worst possible case for predicting CSFM and assigns a weight on the least possible CSFM value that matches the total absolute error [57; 79].

Top cause accuracy is the number of correct cause of death of being first assignment divided by number of causes.

Top 3 cause accuracy is the number of correct cause of death within the first three cause assignment divided by number of causes [57].

Kappa metric makes the assumption that chance prediction is informed by the true test set cause composition. This evaluation metric is dependent on the cause composition of the test dataset. It provides a measure of association taking into account all causes and not a cause specific measure of concordance [79].

Case similarity measure is used as performance metric in Case Based Reasoning (CBR). CBR evaluates case similarity by using the nearest neighbor approach because of its simplicity and closeness to human judgement. Features are compared to determine the similarity of all past cases in a case base. Feature similarities and the attribute values of the current case are matched against attribute values of past cases, which must be within the system vocabulary. After the similarity measures, similarity (A,B) for all cases are computed; the system ranks similar cases based on the similarity scores. The k-nearest neighbour case(s) can also be obtained for purposes of comparison. Case(s) which scores the highest value of similarity (A,B) will be used by the system as the most probable solution and will be analysed further to suit the new solution [27].

E. ML Optimisation techniques in VA

ML learning entails applying a learning algorithm where a classifier is induced from training data, hence every classifier has a measure of performance (through bias and variance) known as the prediction error which is unknown and is estimated from data [38; 80].

1) ***k-fold cross validation*** *:* Validation analysis is crucial in ML as it finds optimum models to fit with new observations,thus improving accuracy and model prediction. One approach that has evidently been applied in the VA field as an optimisation technique is the k-fold cross validation. k-fold cross validation is a technique that divides the data into k folds, and a classifier learns through $k-1$ folds where the error is computed by testing the remaining fold and denotes the average value of errors committed in each fold. Wong [81] points out that k-fold is the widely used procedure to test the effectiveness of a classification algorithm by comparing the performance between two algorithms. Moreover, performance can be evaluated through by average k-accuracies from k-fold cross validation. Jeblee et al. [3] used 10-fold cross validation and deduced the ANN as the best classifier.

F. Imbalanced and fragmented datasets

There is little literature in VA that discusses the issue of fragmented or imbalanced datasets, also known as skewed datasets, even though it is a common situation in VA domain to have these natural datasets due to the rare situation of the events under consideration. VA data has highlighted a high level of sparseness and imbalance from the text and also in the cause of death categories which is an issue for NLP [82]. Mujtaba et al. [25] points out that imbalanced class distribution (majority class out numbers minority class in terms of records) is when one class can be of more interest and is insufficiently represented compared to other class. In such instances, when categorisation is performed there can be misinterpretations and misclassification. Henceforth, imbalanced datasets are a drawback in algorithmic predictions as there is a always a bias of rules that concentrate mainly within the majority class and tend to frequent less the minority class. This leads to misinterpretation of the results in terms of accuracy in the minority class even though overall performance maybe high. As such, there is need for special techniques to address this thereof.

One approach that has been proposed is over-sampling. This approach increases the size of the minority class through random replication of positive instances, thus a balanced distribution is sought without adding any new information to the dataset. However, over sampling can suffer from issues of model over-fitting because of replicating existing positive cases. Therefore to address over-fitting, an alternative of over-sampling technique called SMOTE (Synthetic Minority Over-sampling Technique) can be applied.

SMOTE uses data generation, such that synthetic minority examples are created by interpolating between pre-existing positive instances that lie close together. Another approach is under-sampling which randomly removes the negative cases from the dataset to make the dataset balanced [25]. This under-sampling can lead to loss of valuable information from the majority class. On the other hand, random over-sampling compensates for the loss by duplicating the cases from the minority class, which may lead to over-fitting. However, despite a plethora of benefits such as ease of implementation and improved accuracy levels, these approaches are worse in terms of performance as compared to other techniques [14]

Kocheturov et al. [14] discusses other approaches of overcoming this data imbalance, which are random sampling, applying the k-nn and choosing an appropriate way for classification quality using techniques such as Area Under Curve or ROC can prove beneficial. Sampling techniques are the most intuitive solution to the problem of imbalanced data. Both nearest neighbours and points on the line segments are chosen at random. Other sampling techniques which integrate machine learning approaches is the Balance Cascade ensemble algorithm that removes previously correctly classified samples from the majority class and randomly samples from the remaining cases of it to construct a new classifier. The sampling technique can be combined with the concept of k-nn. For instance, the NearMiss algorithm selects a number of the closest majority data points for each minority example to guarantee that every minority example is surrounded by some majority ones. The following extract discusses limitations in statistical and ML models.

G. Model limitations and imposed assumptions

The VA community has witnessed the application of statistical approaches such as the NBC and logistic regression. The NBC assumes that features are conditionally independent and normally distributed numeric values. However, the proposition of normal distribution in the NBC does not always hold in that there might be need to approximate other continuous distributions. Performance of this model is very poor when features are highly correlated and the feature selection process can compromise the model [24]. NBC assumes conditional independence of extracted features, however this proposition becomes more complex with the increasing number of features, which ultimately reduces the performance of NBC [18]. Moreover, if there is an imbalanced dataset in terms of training and test data, the NBC can lead to erroneous results and misinterpretations [21]. Consequently, the NBC performance is affected negatively in instances where, NBC assumes conditional independence among features which might not be the case for that specific dataset. Furthermore, conditional dependence in features becomes more complex as the number of features increases [19].

Another statistical technique in VA is Logistic Regression (LR). The model assumes a linear relationship between the inputs and the log odds. However, in reality the proposition of linearity does not always hold true, hence we can have a non-linearity existence in the LR. To overcome the non-linearity problem in LR we recommend the SVM, which uses the kernel trick and is efficient in non-linear classification techniques [14]. As such, SVM tends to be computationally expensive when applying the kernel technique during learning. Moreover, achieving real space division is difficult or rather impossible, thus the SVM applies a margin that allows misclassification of some examples whilst increasing the overall performance. In such cases in biomedicine we use the soft margin SVM through introducing penalties to the objective function [14].

Further inquiries in literature propose the use of a non-parametric technique, k-nearest neighbour method in VA scoring. However, the k-nn method is computationally expensive as there is need to calculate a weighting distance for each data record stored during classification. Mujtaba et al. [19] reported worst results of the k-nn approach citing that this model does not learn from a training set and instead utilizes the training set itself for classification, an example of over-fitting. This further implies that k-nn does not generalize the classification problem effectively and is not robust for noisy data. Moreover, to predict CoD for new autopsy cases, this model identifies the k-nn(s) to the new instance from the training set, and the predicted class label will be assigned as the most common label in the k-nn approach.

The application of the RF in VA has witnessed issues in terms of performance, specifically when the dataset contains large volume of features and very few discriminative features. This therefore, leads to trees in a forest populated by less relevant features that lead to misinterpretations and incorrect predictions [19]. Mujtaba et al. [19] explored the J48 decision tree to classify cause of death from forensic autopsy reports and pointed out that, the model achieved the lowest performance. This was due to issues of having continuous data representing all features in the master feature vector, thus this limited finding the optimal thresholds (inability to perform global optimization because of its local search strategy) needed to construct the J48 decision tree. They concluded that the J48 decision tree is unsuitable for classifying forensic autopsy reports. On contrary the ANNs exhibit better performance because of their ability to handle complex data, when compared to other techniques in terms of accuracy, however it lacks interpretability. Boulle et al. [32] argues that the ANN has limitation of input selection not being straight forward, time consuming in terms of designing optimal network for each cause of death, generalisation of predictions has low sensitivity and specificity and comparing and prioritising CSFM over sensitivity and specificity is a manual process.

Deep Learning models have not been applied extensively in VA, even though deep learning models suffer from interpretability they show promising results. Mujtaba et al. [25] argues that, although CNN showed considerable classification accuracy, its major limitation is the amount of data that it needs to learn the complex features from a clinical report dataset. Therefore, CNN will show limited classification accuracy with only a few instances in the training set. The reason is that CNN has to learn several feature weights to determine the most discriminative and result-oriented features for classifying the clinical reports. Thus, enormous training data are required to achieve this objective. In the clinical text classification domain, the clinical reports related to any particular disease may be generally unavailable in large volumes. In such cases, other alternatives of pre-trained deep learning models (I-H) can be

maximized to classify clinical reports compared with CNN. The following section discusses emerging trends in deep learning.

H. Emerging Trends: Deep Learning in VA

Machine learning techniques suffer from various drawbacks, hence the need to use other emerging technologies of deep learning. Miotto et al. [43] argues that there are various challenges with traditional ML techniques, such as supervised ML requiring a labelled dataset by experts, going through data cleaning, performing feature engineering among many. This therefore, leads to poor scaling of the feature space and consequently hides possible opportunities to discover new trends. Boulle et al. [32] argues that conventional ML approaches are limited in their ability to process natural data in raw form because of the composition of single, usually linear transformation of the input space. As an option, deep learning architectures provide representation techniques of the target value from the raw text, thus discovering new patterns and trends.

Deep learning entails various layers made up of neural networks that process data to learn hierarchical abstract representations of data. Deep learning involves NLP, which uses a variety of computational methods for automated analysis and illustration of the human language [66]. NLP tasks done by computers include among many; parsing, tagging part-of-speech, machine translation, computer vision, pattern recognition, semantic role labelling, named-entity recognition, deep learning methods and dialogue systems [67]. The following extracts will discuss other forms of character embeddings that can prove essential in the VA domain.

*1) **Other Recurrent Neural Networks approaches**:* Recurrent Neural Networks (RNNs) is a form of character embedding architecture that is employed in NLP to process sequential information. Therefore apart from the GRNN applied in Yan et al. [20], there is need to explore with other RNNs techniques in VA research such as the Long Short Term Memory (LSTM) architecture. The LSTM technique has the capability to learn and hyper tune parameters for optimum results, thus overcoming the issue of the vanishing gradient. Figure 5 is the RNN model.

Long term short memory

Long Short Term Memory (LSTM) character embedding architecture that has the forget, input and output gate. The forget gate makes it possible for it to error back propagate a number of times and calculates the hidden layer by averaging the three layers [66]. It calculates the hidden state by combining the three gates. Figure 6 is the LSTM model.

Matrix Vector Recursive Neural Network

The Matrix Vector Recursive Neural Network is a character embedding architecture, that seeks to represent every word and phrase as both matrix and vector [83]. It creates a combination of one matrix multiplied by the vector of the other. Recursive Neural Tensor Network (RNTN) applies less parameters and promotes more interaction between input vectors [66].

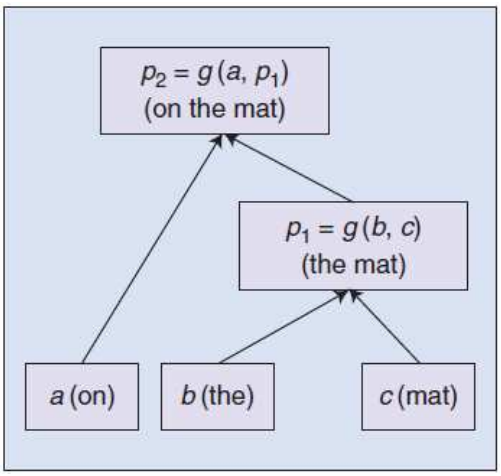

Fig. 5. Recurrent neural network framework.

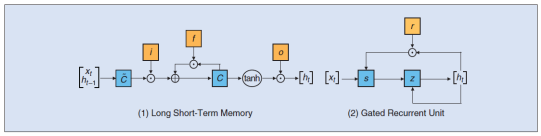

Fig. 6. Long term short memory framework.

Transformer

Developments in deep learning have led to new architectures such as the transformer, which has an encoder that maps input sequences of symbols to a sequence of continuous representations, such that the decoder generates an output sequence of symbols one element at a time. The transformer takes input of previous instances to generate next feature representations [84]. Vaswani et al. [84] points out that the transformer architecture uses a decoder and encoder that are connected layers with structured attention. The encoder takes symbol representations in a sequence $(x_1, ..., x_n)$ to continuous sequenced representations $(z_1, ..., z_n)$, thus decoder uses z sequenced outputs $(y_1, ..., y_m)$ features one at a time. At each stage, the model takes previous input to generate the next symbols. The encoder has 6 stacks of similar layers which have 2 sub layers apiece for attention and feed forward functionality. The decoder is made up of 6 stacked similar layers and 2 sub layers apiece. It then has a third sub layer which functions as multi head attention over output from the encoder. The attention has mappings to requests and uses keys with values that correspond to output. It creates vectors for requests/ queries, keys, values and output. The end result is a calculated weight of all values using a compatibility function of the request and the related key. It applies the scaled dot product that takes input as queries of dimension d_k and values d_v. It applies the softmax function to obtain weight by taking the dot product

of the query and keys and divide each by $\sqrt{d_k}$. The multi-head attention attends to varying representations at differing positions in space by permitting the model in a combined way.

I. Limitations in VA literature

This section details a list of limitations which have not been explicitly considered in VA literature;

- Missing data and varying VA input
- No model transparency performed,
- Data balancing techniques not applied,
- Lack of standardization of tools and methods used in VA studies,
- Scarcity and cost of experts,
- Various medical ontologies resulting in heterogeneous datasets,
- Relying on expert's diagnosis as gold standard,
- Scarce availability of publicly available biomedical datasets with standard training and testing data,
- High dimensional low sample size datasets and noisy data points,
- No exploratory data analysis is done to explore variable distribution and check for any outliers in datasets and
- Model Complexity is not taken into account.

This section synthesises the main limitations of VA literature. In general all the discussed methods take input of VA narratives, however the end result of categorisation is affected by main limitations of VA studies. VA studies mainly employ small sample sizes of studies that limit the precision of estimates, moreover, there is bias in interviewees remembering true cause of deaths, which can result in misclassification. Furthermore, many deaths in VA studies cannot be determined and are either excluded from the study (due to incomplete information) or are classified as unknown cause of death [85].

Most statistical and machine models in VA do not perform model transparency and this reduces the confidence in model interpretability, hence there is need to incorporate model transparency in order to be able to justify and explain model prediction. The VA community can apply transparency models such Local Interpretable Model-Agnostic Explanations (LIME) amongst many to cater for this limitation. Few studies report on data balancing to improve classification accuracy, even though there is scarce literature on the actual application of data balancing techniques in VA studies, one can incorporate SMOTE as a data balancing approach amongst many.

There is also a need to have standardization of tools, processes and methods used in VA studies, one example is the diverse options in data acquisition and preprocessing. Moreover, one key challenge and limitation is the scarcity of physicians that perform VAs. Thus, there is need to prioritise resources in order to address the imbalance. This synthesis of the literature revealed the one main limitation to be heterogeneous datasets. Hence, there is a great need to have standard ontologies that house VA data.

Further inquiries from the literature revealed that there is no gold standard that can be used to test the performance of VA. This lack of a true gold standard limits the explainability and interpretation of validation study findings. To add more to these limitations is the issue of scarce availability of biomedical datasets that can be used to train and test models in VA in order to improve performance. It is key going forward to have more publicly available datasets that can be used by the VA community to train and test models effectively, as this will improve model performance and avail better results.

There is need to have more robust dimensionality reduction schemes in place and therefore, special preprocessing techniques or noise robust algorithms are needed. Although feature reduction/selection improves model accuracy, the literature needs to look at other feature engineering techniques in order to improve model performance and genarate new features, such as genetically inspired feature selection or engineering approaches. Model complexity is a result of models that are highly parameterised, which is the norm in biomedical data, therefore, it is key to keep a few parameters during model development.

II. Methods

This study employs a systematic literature survey approach to;

1) systematically review the most commonly used statistical and machine learning techniques in determining causes of deaths from VAs;
2) identify limitations of each of these techniques;
3) propose a guiding machine learning framework for determining causes of death from VAs;
4) point to emerging directions;

In this survey, the statistical techniques considered include, Logistic Regression (LR) and Naïve Bayes (NB), and the machine learning approaches include k-Nearest Neighbor (k-NN), Decision Trees (DTs), Support Vector Machines (SVMs), Artificial Neural Networks (ANN), Random Forests (RFs), Case Based Reasoning (CBR), Ensemble classifiers (Boosting and Bagging), k-means clustering, distributed word vectors (Word2vec, Sentence2vec, Doc2vec, Glove), Convolutional Neural Network (CNN) and Gated Recurrent Neural Network (GRNN). This list of statistical and machine learning approaches is not exhaustive from the range of techniques that are used to model VAs. This study only details the most commonly applied techniques. A process flowchart of the systematic literature survey methodology is presented in Figure 8. This study follows the methodology of the recently published systematic literature survey articles on VA. The study used the following search terms: "Algorithms in Verbal Autopsy", "Statistical techniques in Verbal Autopsy", "Machine Learning in Verbal Autopsy" and "Deep Learning in Verbal Autopsy" to retrieve relevant articles. The

inclusion criterion included published articles on VA over the period 2010 to 2020. The study selected peer-reviewed journal and conference articles as these are considered to be of high quality. The articles were selected mostly by reading the abstract and the conclusion and the entire article in some in instances. Only articles that were written in English were considered. The repositories used in the search for papers were Google Scholar, PlosOne, Science Direct, IEEEXplore, BioMed Central and Springer-Link. These databases include studies which are undertaken worldwide, hence geographical bias is removed. All unpublished work and dissertations were not included. The search yielded 110 articles of which 76 where peer reviewed journal articles and 34 where conference papers. Furthermore, we had a database that served as a point of reference for further statistical analysis. The main intention was to quantitatively establish the research status of statistical and machine learning application in VA. The literature meta analysis identified 85 articles as being relevant for our study and the remaining 25 were discarded. A meta-analysis of the results from the selected articles is done by producing summary tables, pie charts and histograms. Figure 7 depicts a histogram of the initial peer reviewed journal articles and conference papers.

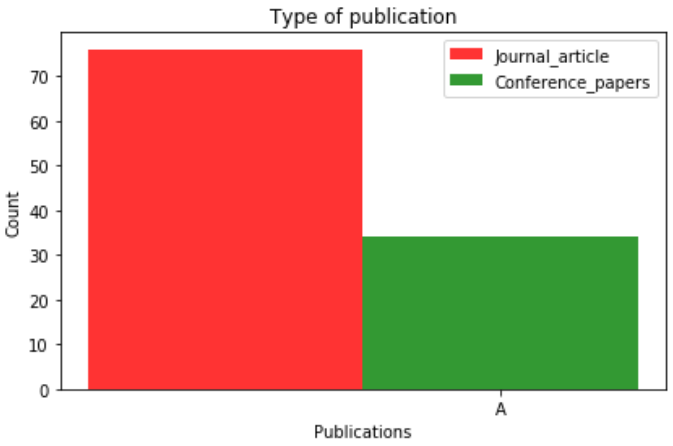

Fig. 7. Distribution of initial searched journal articles vs conference papers

A. *Existing literature survey on VA*

There are few literature surveys on VA [5; 8; 25; 28]. In this study we briefly highlight the objectives of each literature survey and show where they differ with our study. The literature surveys focus mostly on comparing models determination of causes of deaths from VAs based on their performances. However, the listed literature surveys ignore key factors in determination of causes of deaths from VAs, such as pre-processing and data exploration, model explainability (transparency) and nature of datasets. Furthermore, deep learning models are not considered in the listed table below of literature surveys. We have also included a study done by Mujtaba et al. [25], which provides the most recent and the most comprehensive comparative analysis of existing methods in automated text classification in VA. The included studies cover mainly performance evaluation of traditional techniques and propose new methods for VA classification. The literature discussed in Mujtaba et al. [25] covers most automated text classification techniques in VA, model performance, propose new methods for VA, feature selection, feature engineering and other issues. However, in contrast to this study, there has not been a systematic literature survey for VA that has been published which encompasses, statistical learning, traditional machine learning and emerging deep learning models. Table I provides a comparison of existing literature survey papers in VA against this study.

III. Results

This section gives an account of results from the primary studies, based on the meta-analysis. The frequency distributions of, feature extraction techniques, feature value representation schemes, feature selection approaches, models and evaluation metrics from all primary studies are analyzed. The pie-charts for pre-processing techniques, studies that report on data transparency and balancing techniques, the distribution of selected VA articles compared to other domains are also shown. We present a tag cloud of all articles used in this review.

A. *Preprocessing*

Pre-processing is a key stage in cleaning data. Our review suggests that only 38% of the primary studies applied pre-processing to normalise their data. There is 62% of primary studies that did not apply pre-processing to their data. As such, this adds to noise and high dimensional data and consequently low accuracy prediction models. Figure 9 below denotes the distribution of studies that applied pre-processing.

B. *Transparency*

Transparency is an approach that makes machine learning models more explainable and justified. As such, there is not a single study in VA that applies transparency models and we only present 5% of the studies that report on transparency. There is great need to introduce transparency in VA models. Figure 10 below shows the studies that discuss transparency models even though they where not specifically applied.

C. *Data balancing*

VA data generally entails data imbalance where there is more bias in terms of the majority class. Approaches such as oversampling and undersampling have been applied to address this imbalance. However, few notable studies apply such approaches. Figure 11 shows only 6% of the primary studies applied data balancing.

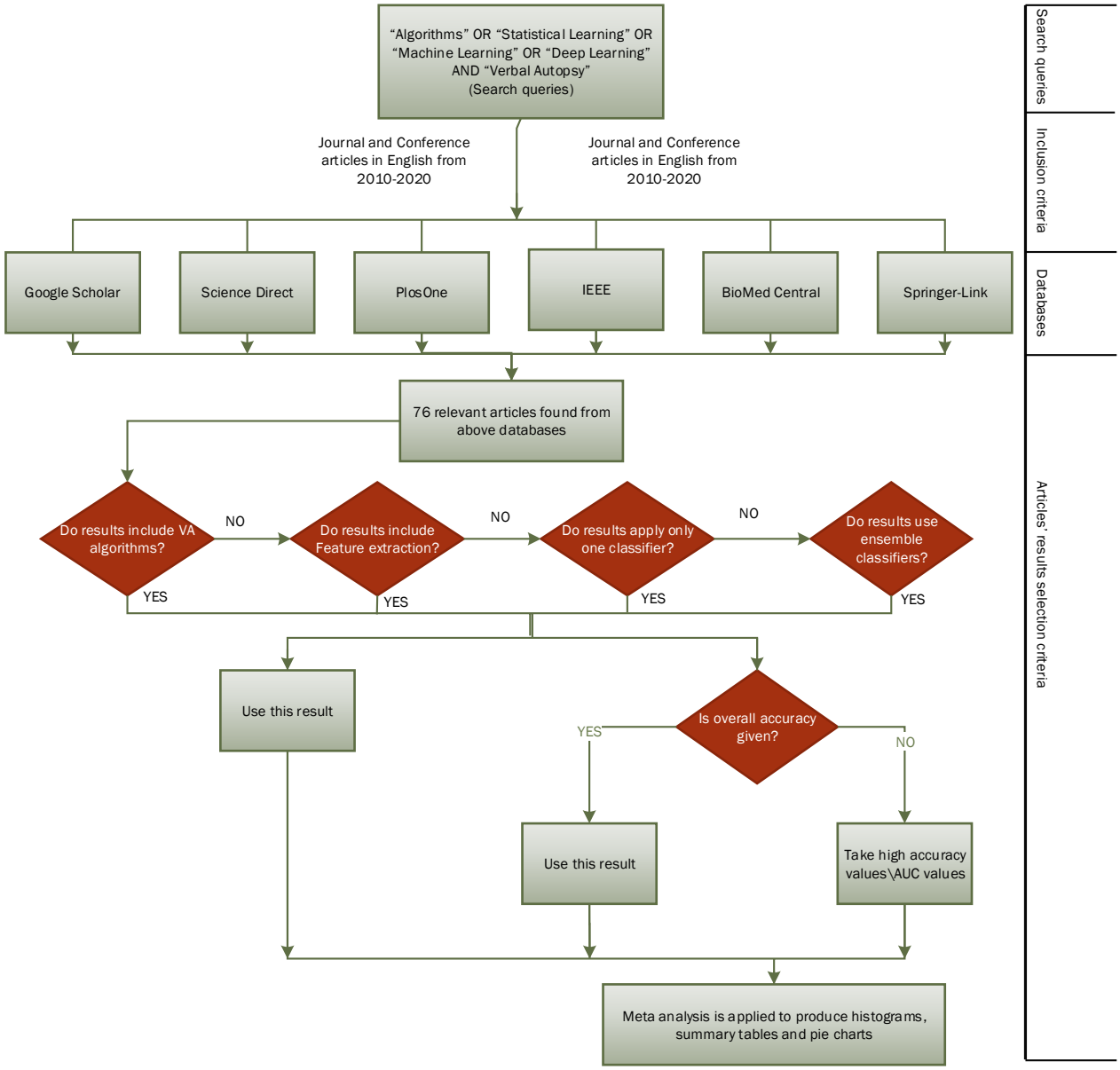

Fig. 8. Process flowchart of the systematic literature survey methodology.

D. Feature extraction

Table II and Figure 12 details feature extraction techniques from primary studies. Our analysis shows that out of 38 studies, most of the VA models use expert driven feature selection, followed by the bag of words and n-gram models respectively.

E. Feature value representation schemes

Table III and Figure 13 shows the distribution of feature value representation schemes. This investigation suggests that out of the 11 studies analysed, term frequency value representation and binary representation are the most widely used approaches.

TABLE I
EXISTING LITERATURE SURVEYS ON VA AND THEIR DIFFERENCES FROM THIS SURVEY PAPER: LEGEND:FEATURE ENGINEERING (FE), FEATURE SELECTION (FS).

Comparative analysis of existing literature papers				
Survey paper	Articles searched	Years	Objectives	Differences from this study
[28]	Unspecified	1997-2009	Model performance evaluation	Deep learning models not covered
			Investigation of model types	Nature of datasets not reported
			Analysis of FE approaches	Not all ML approaches are discussed
				Model transparency not covered
[25]	1497	2013–2018	Model performance evaluation	Deep learning not covered
			Investigation of model types	
			Analysis of pre-processing approaches	Not all ML approaches are discussed
			Analysis of FE approaches	
			Analysis of performance metrics	
[5]	19	1997–2014	Model performance evaluation	Deep learning not covered
			Investigation of model types	
				Not all ML approaches are discussed
			Analysis of FE approaches	
			Analysis of performance metrics	
				Model transparency not covered
[8]	392	1978–2008	Model performance evaluation	Deep learning not covered
			Comparative analysis of VA methods and concepts	Nature of datasets not reported
				Not all ML approaches are discussed
				Model transparency not covered

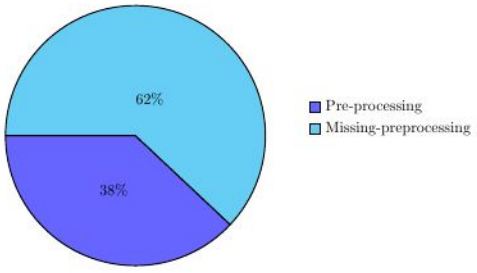

Fig. 9. Distribution of VA primary studies that include pre-processing.

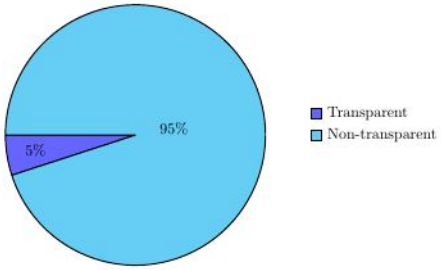

Fig. 10. Distribution of VA primary studies that include transparency.

F. Feature selection

Table IV and Figure 14 shows the distribution of feature selection techniques applied in primary studies. Out of 18 studies, NLP is the most common approach possibly depicting deep learning as an emerging solution, Pearson Correlation and Information Gain coming after.

TABLE II
FEATURE EXTRACTION METHODS USED IN PRIMARY STUDIES. LEGEND: BAG OF WORDS (BOW), N-GRAMS (NG), HYBRID (HB), WORD2VEC (WV), CONCEPT BASED (CB), GRAPH BASED (GB), RULE BASE (RB), DOC2VEC (DV), EXPERT DRIVEN (ED), SENTENCE2VEC (SV), GLOVE (GV).

	Feature extraction methods from primary studies										
Study	BoW	HB	NG	WV	CB	GB	RB	DV	SV	GV	ED
[3]	✓										
[4]											✓
[5]											✓
[56]											✓
[6]											✓
[18]		✓	✓								
[28]	✓										
[19]		✓	✓								
[31]											✓
[21]											✓
[22]	✓	✓	✓	✓	✓	✓	✓	✓			
[7]											✓
[29]											✓
[8]											✓
[57]											✓
[58]											✓
[59]											✓
[60]											✓
[61]											✓
[62]											✓
[27]							✓				
[25]	✓	✓	✓	✓	✓	✓	✓				
[18]	✓								✓		
[30]											✓
[32]											✓
[34]											✓
[35]											✓
[42]											✓
[45]	✓										
[46]				✓							
[20]				✓						✓	
[48]			✓		✓						
[54]					✓						
[63]	✓										
[64]	✓										
[69]											✓
[79]											✓
[82]											✓

G. *Comparative analysis of studies reviewed in this paper*

Figure 15 shows the VA study distribution compared to other fields. The distribution shows that we used 66% of the primary studies which are in the VA domain and the remainder where from other fields. Figure 16 below depicts the journal articles and conference papers used in this review. From the primary studies synthesised, we used 66% from reputable and peer reviewed journal articles and 34% from conference proceedings.

H. *Evaluation metrics*

The section below discusses various performance evaluation metrics used in this systematic review. Table V and Figure 17 show the distribution of performance evaluation approaches applied in this review. Out of 34 studies, the most used evaluation metrics are Precision and Recall followed by CSFM accuracy, CCC, Specificity and F- score.

I. *Tag cloud of common and frequent terms of reviewed studies*

Figure 18 is a visual model that illustrates the highest frequency terms appearing in the title and abstract of the peer

TABLE III
FEATURE VALUE REPRESENTATION SCHEMES FROM PRIMARY STUDIES. LEGEND: BINARY REPRESENTATION (BR), TERM FREQUENCY (TF), DOCUMENT FREQUENCY (DF), TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF), NORMALISED TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (NTF-IDF).

Feature value representation from primary studies				
Study	BR	TF	TF-IDF	NTF-IDF
[3]		✓		
[18]	✓	✓	✓	✓
[28]	✓	✓	✓	✓
[22]			✓	
[25]	✓	✓	✓	✓
[45]		✓		
[48]	✓	✓		
[29]		✓		
[54]	✓	✓		
[63]				✓
[82]		✓		

TABLE IV
FEATURE SELECTION TECHNIQUES FROM PRIMARY STUDIES. LEGEND: PEARSON CORRELATION (PC), CHI-SQUARED (CS), INFORMATION GAIN (IG), PRINCIPAL COMPONENT ANALYSIS (PCA),FISHER MARKOV SELECTOR (FMS), IMPROVED GLOBAL FEATURE SELECTION (IGFS), LOCAL SEMI SUPERVISED (LSS), NATURAL LANGUAGE PROCESSING (NLP)

Feature selection techniques from primary studies								
Study	PC	CS	IG	PCA	FMS	IGFS	LSS	NLP
[3]	✓							
[28]							✓	
[57]	✓							
[62]		✓						
[19]	✓	✓	✓					
[62]			✓					
[29]	✓	✓	✓		✓	✓		
[25]	✓	✓	✓	✓				
[18]	✓	✓	✓		✓	✓		
[45]			✓					
[46]								✓
[20]								✓
[54]								✓
[63]								✓
[64]								✓
[82]								✓
[85]								✓
[40]		✓						

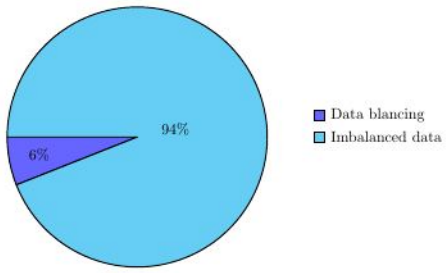

Fig. 11. Distribution of VA related studies that include data balancing.

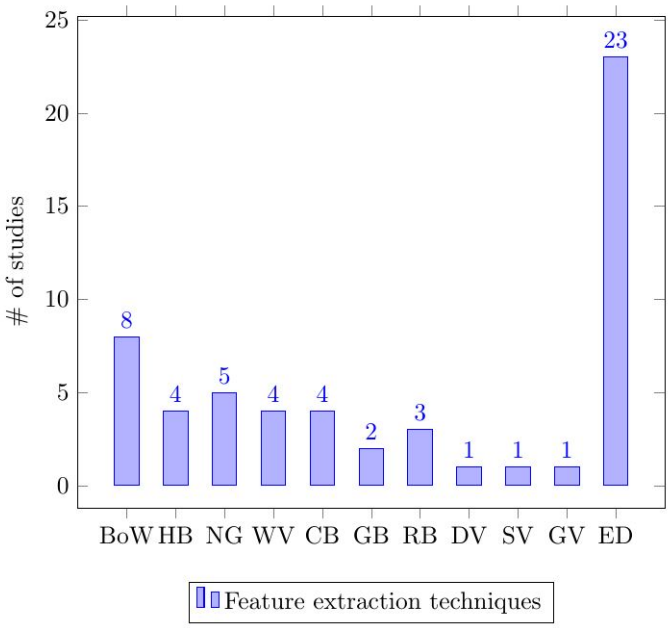

Fig. 12. Distribution of feature extraction techniques reviewed.

reviewed literature, where higher occurring terms appear in larger font size. From our cloud tag we can deduce that ML task of classification appears more, also we can can report that verbal autopsy is common within the health field through text or data that forms part of our narratives.

J. Models

The section below discusses various models in this systematic review. Table VI and Figure 19 show the distribution of statistical and machine learning models applied in this study. This literature review suggests that out of 36 studies, conventional VA algorithms are the most widely used technique, followed by Naive Bayes Classifier, RF and SVM and k-nn.

TABLE V
EVALUATION METRICS TECHNIQUES APPLIED IN PRIMARY STUDIES. LEGEND: SPECIFICITY (SP), PRECISION (PR), RECALL (RE), F-SCORE (FS), ACCURACY (AC), CSFM (CS), CCC, PCCC, KAPPA (KA), CASE SIMILARITY MEASURE (CSM), TOP CAUSE (TC), TOP3COD (T3), AREA UNDER CURVE (AUC)

	Evaluation metrics of primary studies												
Study	Sp	Pr	Re	Ac	CSFM	CCC	PCCC	Ka	CSM	Fs	TC	T3	AUC
[3]		✓	✓		✓		✓						
[4]	✓	✓	✓		✓								
[1]					✓								
[5]	✓	✓			✓	✓							
[6]		✓											
[18]		✓	✓										
[28]		✓	✓							✓			
[19]		✓	✓	✓						✓			
[31]					✓	✓							
[21]	✓	✓			✓		✓						
[22]		✓	✓	✓						✓			
[7]					✓						✓		
[8]	✓	✓	✓		✓			✓					
[57]					✓	✓					✓	✓	
[58]					✓	✓							
[59]					✓	✓							
[60]		✓			✓		✓						
[61]					✓	✓		✓					
[62]	✓			✓		✓							
[25]		✓	✓	✓					✓			✓	
[29]		✓	✓	✓						✓			✓
[30]	✓	✓	✓		✓	✓		✓					
[32]	✓		✓	✓		✓							
[35]						✓		✓					
[42]			✓				✓		✓				
[45]				✓									
[46]		✓	✓							✓			
[20]		✓	✓		✓					✓			
[48]		✓	✓							✓			
[54]			✓	✓							✓		
[63]		✓	✓							✓			
[69]			✓										
[79]	✓	✓	✓		✓	✓		✓					
[85]		✓	✓										
[40]													✓

K. Guiding machine learning framework for VA

Our proposed guiding framework for VA researchers and practitioners is based on the results of this literature survey. First and foremost, because of the nature of datasets in VA as being balanced or imbalanced, we propose performing exploratory data analysis and data pre-processing as first steps of data cleaning. After successful data cleaning, we suggest performing well structured and logical feature selection and engineering to attain the most discriminative and representative features for our vector space. To achieve a balanced dataset we propose doing sampling using the SMOTE approach. In this framework we propose using the Naive Bayes model and random forest classifiers as benchmarks since all current VA algorithms employ these statistical and ML approaches. However, we also recommending using the SVM approach that can handle non-linearity of data and can explain predictions more clearer. This literature survey shows that the SVM and ensemble classifiers perform better as compared to other models. After this process we suggest applying evaluation metrics for performance evaluation of the approaches to choose the optimum one for classification. This systematic survey showed that the most common feature

TABLE VI
MODELS APPLIED IN PRIMARY STUDIES. LEGEND: VA ALGORITHMS (VA), LOGISTIC REGRESSION (LR), RANDOM FORESTS (RF), DECISION TREE (DT), K- NEAREST NEIGHBOR (K-NN), SUPPORT VECTOR MACHINE (SVM), ARTIFICIAL NEURAL NETWORK (ANN), NAIVE BAYES CLASSIFIER (NBC), K-MEANS CLUSTERING (K-MEANS), RULE BASED (RB), ENSEMBLE CLASSIFIER (EC), CONVOLUTIONAL NEURAL NETWORK (CNN), GATED RECURRENT NEURAL NETWORK (GRN)

	Models applied in of primary studies												
Study	VA	LR	RF	DT	k-nn	SVM	ANN	NBC	k-means	RB	EC	CNN	GRN
[3]		✓	✓	✓		✓	✓	✓					
[1]	✓												
[5]	✓		✓										
[56]	✓												
[6]						✓					✓		
[18]			✓			✓		✓					
[28]			✓			✓		✓					
[19]			✓	✓	✓	✓		✓		✓			
[31]	✓	✓	✓										
[21]	✓							✓					
[22]			✓	✓	✓	✓		✓			✓		
[7]	✓												
[8]	✓							✓					
[57]	✓												
[58]	✓												
[59]	✓												
[60]	✓		✓										
[27]					✓			✓		✓			
[61]	✓												
[62]	✓												
[25]		✓	✓	✓	✓	✓	✓	✓				✓	
[29]			✓	✓	✓	✓		✓					
[30]	✓												
[32]		✓					✓						
[34]		✓											
[35]	✓												
[42]								✓			✓		
[45]		✓		✓		✓		✓					
[46]									✓			✓	✓
[20]													✓
[48]					✓				✓	✓			
[54]					✓				✓	✓			
[63]						✓							
[69]	✓												
[79]	✓												
[82]	✓												
[40]		✓											

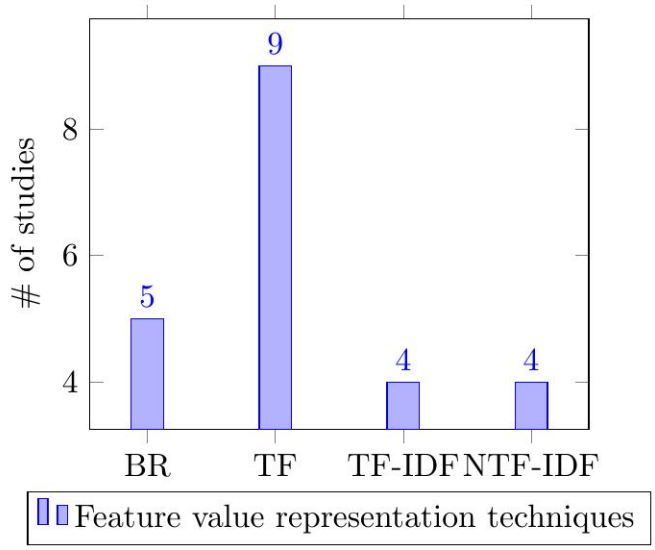

Fig. 13. Distribution of feature value representation schemes.

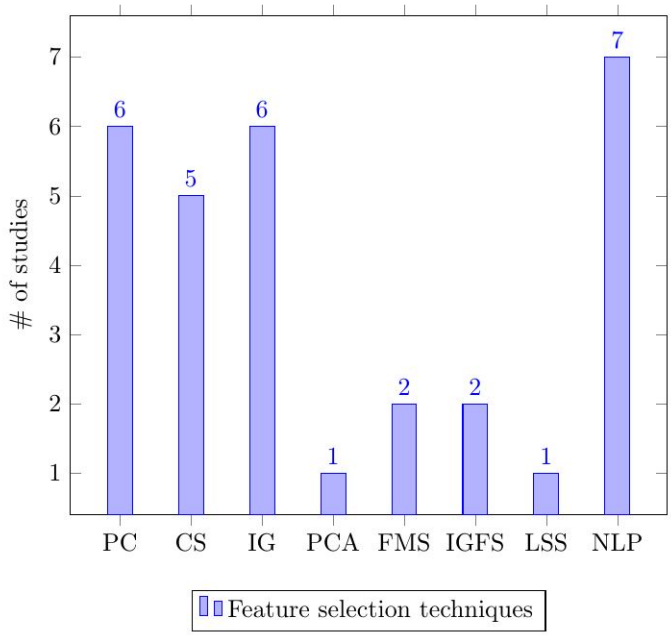

Fig. 14. Distribution of feature selection techniques.

extraction techniques are bag of words and n-grams. The most frequently used feature value representation schemes are BR and TF. Our investigation further showed that the most common feature selection or dimensionality reduction techniques are; expert driven feature selection by physicians, Information gain and Pearson correlation.

Moreover, the few studies that have used the emerging trend of advanced machine learning known as deep learning have proved that distributed vector representations (Word2vec, Sentence2vec, Doc2vec and Glove) show promising results. CNN improves VA performance through use of character

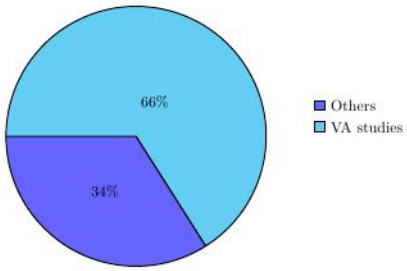

Fig. 15. Distribution of VA related studies in comparison to other fields used in this review.

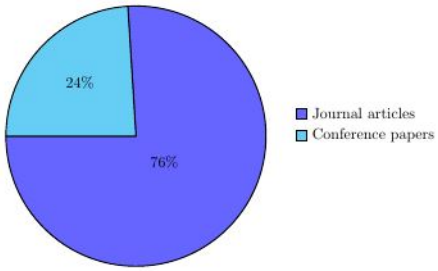

Fig. 16. Distribution of journal articles and conference articles used in this review.

embeddings. New models exist in deep learning such as the Recursive Neural Network and Transformer, which have shown promising results in other domains and can prove useful to explore with, in our VA field. Performance evaluation metrics such as Precision, Recall, Accuracy, AUC and F-score can be applied to evaluate models.
Below Figure 20 is the proposed visual representation of the guiding framework.

IV. Discussion

Our study conducted a comprehensive systematic analysis of publications that are related to the VA field in terms of statistical and machine learning application to determine cause of death. Subsequently, several main sub fields related to ML in VA were presented. We further gave an insight on the application of statistical, ML and advanced deep learning approaches in VA, which has not been done at this stage. Therefore from this detailed meta analysis, one can be able to identify various statistical and ML applications in VA and thus also identify opportunities and challenges in the VA

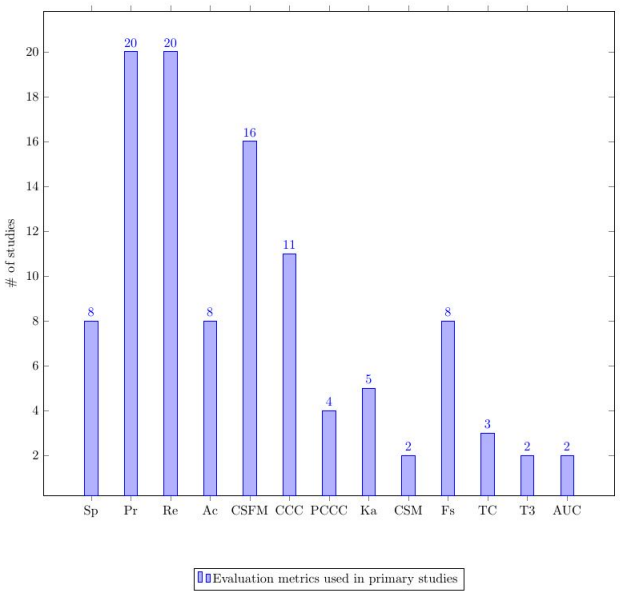

Fig. 17. Distribution of evaluation metrics from primary studies reviewed.

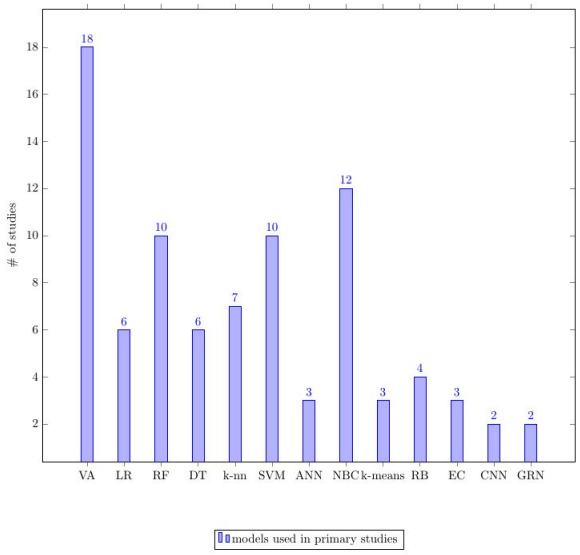

Fig. 19. Distribution of statistical and ML models used in primary studies.

Fig. 18. Tag cloud of this systematic review

domain. Most of the studies carried out in VA have been using mostly supervised learning techniques with a few reports on unsupervised learning approaches. Moreover, there is little or few studies that apply deep learning as a ML approach in VA. Deep learning model application can improve results within the field of VA, such as better knowledge representation systems for VA, thus enforcing standardisation and a shared data repository for further research.

Generally all ML studies in VA have been trained on sparse high dimensional data and high training cost of samples, thus applying alternative approaches can effectively handle increased training data samples and improve efficiency. There is also a limitation in terms of having publicly available training datasets, which can be used by researchers to train and improve models. This should be one key area to be looked into within the VA community.

Future VA studies should incorporate and utilise well structured and logical feature selection and feature engineering techniques such as filter, wrapper and embedded methods, as they have shown promising and excellent results in other fields. These methods will make it possible to attain relevant features that are discriminative, thereby addressing issues of noise and high hyper parameter space amongst many drawbacks [63].

Literature to date has presented various studies that have applied performance evaluation of different statistical and classical machine learning models in VA. However, there has not been any agreement in terms of the best optimum performing model(s). In this study we systematically reviewed and analysed the most commonly used statistical, classical machine learning and deep learning models in VA . The results of this investigation showed that the SVM as the most optimum classifier and we deduced that ensemble classifiers generally outperform single classifiers. However, we have witnessed minimal application of deep learning models in VA, even though the application of the CNN from literature showed improvements to classification accuracy as well as better knowledge representation using character embeddings in the VA domain. This was better in comparison to conventional statistical and classical machine learning models.

Our study further flags limitations such as lack of standardised VA questionnaires, scarcity and cost of experts to perform VAs, various medical ontologies resulting in heterogeneous datasets, relying on expert's diagnosis as the gold standard, scarce availability of publicly available biomedical datasets, diversity of options in data acquisition and preprocessing, high dimension low sample size datasets amongst many. The proposed guiding framework for VA practitioners and researchers in this study suggests performing robust and logical exploratory data analysis and pre-processing for data

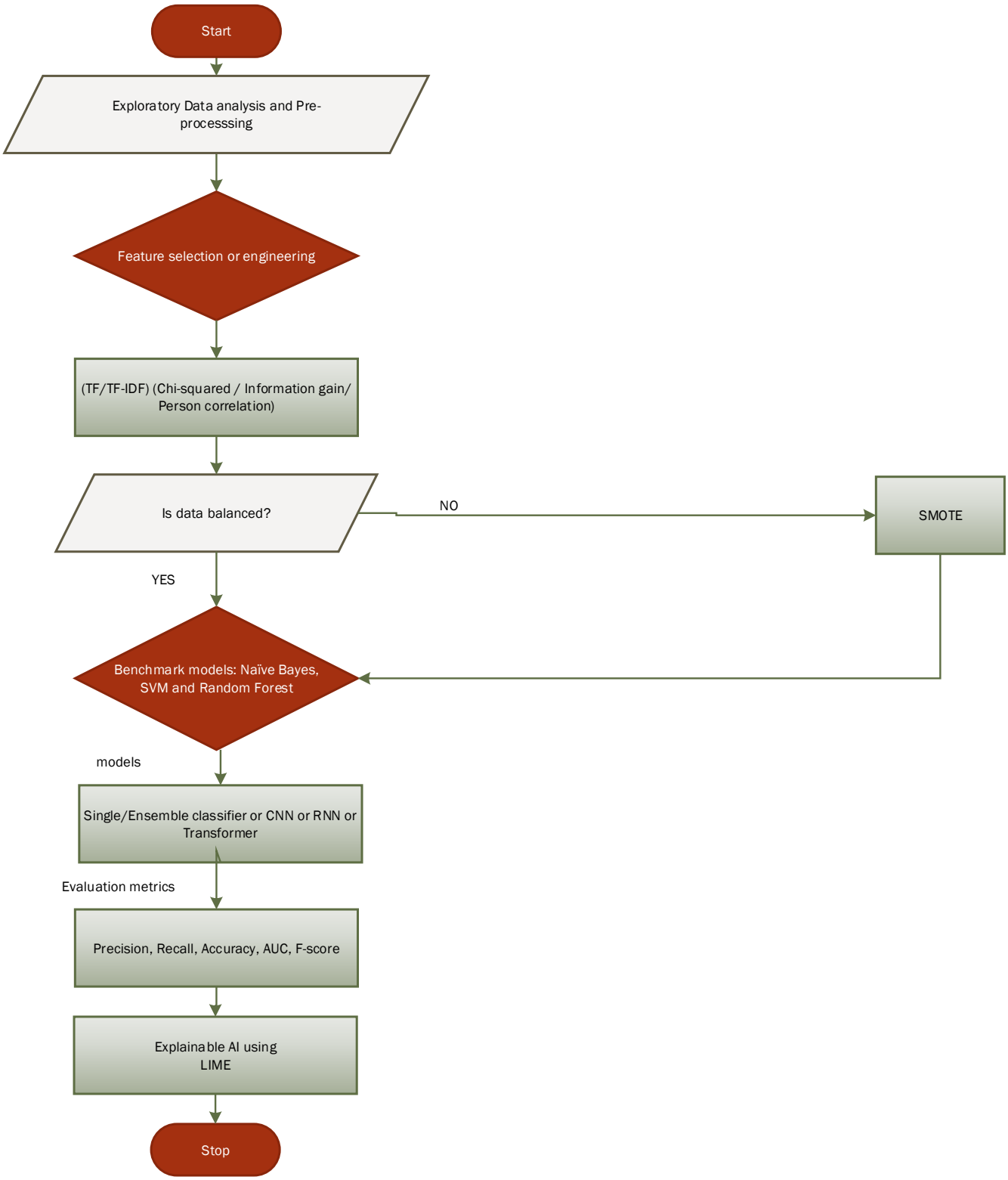

Fig. 20. The guiding machine learning VA that is proposed in this literature survey

cleaning, applying feature selection and feature engineering in the most effective way, performing data balancing, employing benchmark models and applying evaluation metrics for model performance. This survey pointed to emerging directions in VA such as the applications of deep learning models in VA and explainable AI.

Future research should focus more on balancing classes of datasets in VA. Balancing techniques that neither do over-sampling nor under-sampling such as linear dependence approach and wavelet data transformation should be explored in VA. Future research should introduce and apply standard VA questionnaires which are key input to the VA process. Moreover, there should be standardisation of medical ontologies and incorporation of homogeneous datasets. Data acquisition and pre-processing techniques should be standardised and be performed in a robust and logical manner. The literature does not take into consideration the exploratory data analysis, hence future research should focus on this aspect to better understand for example the distributions of features. Future studies should focus more on model complexity to allow efficiency in model development in VAs. Future research should determine the correlation between the target variable and the independent variables/features in-order to identify predictive variables/features. Furthermore, there is need to make available public datasets in order to train models for improved performance. If resources permit, there should be a team of experts that specifically deal with VAs.

The future of the VA community should focus on transparency, to justify and make non-transparent models explainable. One example of a transparent model is Local Interpretable Model-Agnostic Explanations (LIME). LIME can explain any classifier prediction through learning an interpretable model around a prediction. Transparency is key in that it builds trust in humans in terms of predictions, such that if there is need for interventions or any form of action then it can be taken. Transparency gives the leverage for an expert to accept or reject a model.

Most VA studies have employed supervised machine learning approaches. There is need for the VA research community to consider other forms of machine learning such as; unsupervised learning (which do not need a labelled dataset, hence performs classification or clustering based on unlabelled points) , semi-supervised machine learning (makes use of few labelled data points to make predictions), transfer learning (relevant where the input data set has few labelled features to effectively train a model) and reinforcement learning (learns through past experience and interaction with the environments to accurately classify VA narratives through a reward system).
To provide more robust VA estimates moving forward, it is pertinent that the global VA community come together to review the limitations of key VA data sources and methods and define a strategy for how these methods could be improved upon.

V. Conclusion

In this study, we reviewed in detail traditional statistical and machine learning approaches to determine cause of deaths from VAs, identified limitations of each of these techniques, proposed a guiding machine learning framework for determining causes of death from VAs and pointed to emerging directions. Most VA models use the PCVA approach as a gold standard, even though it is expensive, inconsistent and inaccurate (subjective and prone to errors because of the VA narratives that are taken as input). In practice, as an alternative to PCVA, mostly statistical approaches such as LR and NB are employed because of simplicity and transparency. Nevertheless, in literature complex machine learning and emerging deep learning frameworks exist, that can be applied to improve cause of death classification. Most of the machine learning models reviewed apply the RF and SVM. However, incorporating ensemble classifiers can yield better results. To achieve optimum predictions in determining cause of deaths from VAs, models need to employ effective data exploration (fully understand datasets before modelling), data balancing, pre-processing (improve knowledge representation of medical datasets) and robust emerging feature engineering techniques (extract relevant features and narrow down the feature space, thus improving model performance). Additionally, other types of machine learning like transfer learning and reinforcement learning need to be explored further as they have shown promising results in other domains. Since VAs are key inputs to the civil registration, there is need to build transparent models for cause of death classification.

The significant contribution of VA data in informing interventions, health priorities, monitoring and evaluation programmes, is of importance. Therefore, it is imperative for the VA community to continue exploring and developing new strategies in line with the long term efforts of improving the civil registration systems. Advances in VA methods should seek to achieve consistency, comparability, and generalisation to other studies.

VI. Abbreviations

The list of abbreviations is given in Table VII below.

VII. Declarations

A. *Ethics approval and consent to participate*

This is not applicable to this study.

B. *Consent for publication*

This is not applicable to this study.

C. *Availability of data and materials*

Data sharing is not applicable to this article as no datasets were generated or analysed during the current study.

D. *Competing interests*

The authors declare that they have no financial or personal relationship(s) that may have inappropriately influenced them in writing this article.

TABLE VII
LIST OF ABBREVIATIONS USED IN THIS STUDY

List of abbreviations	
Abbreviation	Full term
AI	Artificial intelligence
ANN	Artificial neural network
AUC	Area under curve
BoW	Bag of words
BR	Binary representation
CBR	Case based reasoning
CCC	Chance corrected concordance
CCVA	Computer coded verbal autopsy
CNN	Convolutional neural network
CSFM	Cause specific mortality fractions
DT	Decision tree
FE	Feature engineering
FMS	Fisher markov selector
FN	False negative
FP	False positive
FS	Feature selection
GRNN	Gated recurrent neural network
GoW	Graph of words
ICD-10	International classification of diseases, tenth revision codes
IG	Information gain
IGFS	Improved global feature selection
k-nn	k-nearest neighbour
LIME	Local interpretable model-agnostic explanations
LSFS	Local semi-supervised feature selection
LR	Logistic regression
LSTM	Long term short memory
ML	Machine learning
NB	Naive bayes
NLP	Natural language processing
NTF-IDF	Normalised term frequency-inverse document frequency
PCA	Principal component analysis
PC	Pearson correlation
PCVA	Physician certified verbal autopsy
RF	Random forest
RNN	Recurrent neural network
SMOTE	Synthetic Minority over-sampling technique
SVM	Support vector machine
TF	Term frequency
TF-IDF	Term frequency-inverse document frequency
TN	True negative
TP	True positive
VA	Verbal autopsy

E. Declaration of interest

The authors declare that they do not have any financial and personal relationships that could inappropriately influence this study.

F. Funding

This research was not grant funded.

G. Authors' contributions

All authors listed below have read and approved the manuscript.

MM is the main author and conducted the whole review process including the write up.

EM reviewed and contributed on statistical models.

CK reviewed and contributed on verbal autopsy theory.

TC reviewed and contributed on the machine learning concepts.

H. Corresponding authors

Michael T. Mapundu

University of the Witwatersrand

School of Public Health

Department of Biostatistics and Epidemiology

7 York Road, Parktown, Johannesburg, South Africa

Tel: +27 11 71-72627

Email: michael.mapundu@wits.ac.za

I. 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.

REFERENCES

[1] Erin K Nichols, Peter Byass, Daniel Chandramohan, Samuel J Clark, Abraham D Flaxman, Robert Jakob, Jordana Leitao, Nicolas Maire, Chalapati Rao, Ian Riley, et al. The who 2016 verbal autopsy instrument: An international standard suitable for automated analysis by interva, insilicova, and tariff 2.0. *PLoS medicine*, 15(1):e1002486, 2018.

[2] Lisa-Marie Thomas, Lucia D'Ambruoso, and Dina Balabanova. Verbal autopsy in health policy and systems: a literature review. *BMJ global health*, 3(2): e000639, 2018.

[3] Serena Jeblee, Mireille Gomes, Prabhat Jha, Frank Rudzicz, and Graeme Hirst. Automatically determining cause of death from verbal autopsy narratives. *BMC medical informatics and decision making*, 19(1):127, 2019.

[4] Nadia Soleman, Daniel Chandramohan, and Kenji Shibuya. Verbal autopsy: current practices and challenges. *Bulletin of the World Health Organization*, 84:239–245, 2006.

[5] Jordana Leitao, Nikita Desai, Lukasz Aleksandrowicz, Peter Byass, Pierre Miasnikof, Stephen Tollman, Dewan Alam, Ying Lu, Suresh Kumar Rathi, Abhishek Singh, 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*, 12(1):22, 2014.

[6] Sean T Green and Abraham D Flaxman. Machine learning methods for verbal autopsy in developing countries. In *AAAI Spring Symposium: Artificial Intelligence for Development*, 2010.

[7] Tyler H McCormick, Zehang Richard Li, Clara Calvert, Amelia C Crampin, Kathleen Kahn, and Samuel J Clark. Probabilistic cause-of-death assignment using verbal autopsies. *Journal of the American Statistical Association*, 111(515):1036–1049, 2016.

[8] Edward Fottrell and Peter Byass. Verbal autopsy: methods in transition. *Epidemiologic reviews*, 32(1): 38–55, 2010.

[9] Min Chen, Yixue Hao, Kai Hwang, Lu Wang, and Lin Wang. Disease prediction by machine learning over big data from healthcare communities. *IEEE Access*, 5: 8869–8879, 2017.

[10] Sharathkumar Anbu and Bhaskarjit Sarmah. Machine learning approach for predicting womens health risk. In *Advanced Computing and Communication Systems (ICACCS), 2017 4th International Conference on*, pages 1–4. IEEE, 2017.

[11] Victor Mudenda, Sebastian Lucas, Aaron Shibemba, Justin O'grady, Matthew Bates, Nathan Kapata, Samana Schwank, Peter Mwaba, Rifat Atun, Michael Hoelscher, et al. Tuberculosis and tuberculosis/hiv/aids–associated mortality in africa: The urgent need to expand and invest in routine and research autopsies. *Journal of Infectious Diseases*, 205(suppl_2):S340–S346, 2012.

[12] B Nithya and V Ilango. Predictive analytics in health care using machine learning tools and techniques. In *2017 International Conference on Intelligent Computing and Control Systems (ICICCS)*, pages 492–499. IEEE, 2017.

[13] Zahid Iqbal, Rafia Ilyas, Waseem Shahzad, and Irum Inayat. 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.

[14] Anton Kocheturov, Panos M Pardalos, and Athanasia Karakitsiou. Massive datasets and machine learning for computational biomedicine: trends and challenges. *Annals of Operations Research*, pages 1–30, 2018.

[15] Lei Liu. Research on logistic regression algorithm of breast cancer diagnose data by machine learning. In *2018 International Conference on Robots & Intelligent System (ICRIS)*. IEEE, 2018.

[16] Ivo D Dinov, Ben Heavner, Ming Tang, Gustavo Glusman, Kyle Chard, Mike Darcy, Ravi Madduri, Judy Pa, Cathie Spino, Carl Kesselman, et al. Predictive big data analytics: a study of parkinson's disease using large, complex, heterogeneous, incongruent, multi-source and incomplete observations. *PloS one*, 11 (8):e0157077, 2016.

[17] Abraham D Flaxman and Theo Vos. Machine learning in population health: Opportunities and threats. *PLoS medicine*, 15(11), 2018.

[18] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, and Khairunisa Shaikh. Automatic text classification of icd-10 related cod from complex and free text forensic autopsy reports. In *Machine Learning and Applications (ICMLA), 2016 15th IEEE International Conference on*, pages 1055–1058. IEEE, 2016.

[19] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, and Khairunisa Shaikh. Prediction of cause of death from forensic autopsy reports using text classification techniques: A comparative study. *Journal of forensic and legal medicine*, 57:41–50, 2018.

[20] Zhaodong Yan, Serena Jeblee, and Graeme Hirst. Can character embeddings improve cause-of-death classification for verbal autopsy narratives? In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 234–239, 2019.

[21] Pierre Miasnikof, Vasily Giannakeas, Mireille Gomes, Lukasz Aleksandrowicz, Alexander Y Shestopaloff, Dewan Alam, Stephen Tollman, Akram Samarikhalaj, and Prabhat Jha. Naive bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC medicine*, 13 (1):286, 2015.

[22] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, Khairunisa Shaikh, and Mohammed Ali Al-Garadi. Classification of forensic autopsy reports through conceptual graph-based

document representation model. *Journal of biomedical informatics*, 82:88–105, 2018.

[23] Suresh Kumar and Shivani Goel. Enhancing text classification by stochastic optimization method and support vector machine. *International Journal of Computer Science and Information Technologies*, 6(4), 2015.

[24] Vandana Korde and C Namrata Mahender. Text classification and classifiers: A survey. *International Journal of Artificial Intelligence & Applications*, 3(2): 85, 2012.

[25] Ghulam Mujtaba, Liyana Shuib, Norisma Idris, Wai Lam Hoo, Ram Gopal Raj, Kamran Khowaja, Khairunisa Shaikh, and Henry Friday Nweke. Clinical text classification research trends: systematic literature review and open issues. *Expert Systems with Applications*, 116:494–520, 2019.

[26] Vianney Jouhet, Georges Defossez, Anita Burgun, Pierre Le Beux, P Levillain, Pierre Ingrand, Vincent Claveau, et al. Automated classification of free-text pathology reports for registration of incident cases of cancer. *Methods of information in medicine*, 51(3):242, 2012.

[27] Wei Liang Yeow, Rohana Mahmud, and Ram Gopal Raj. An application of case-based reasoning with machine learning for forensic autopsy. *Expert Systems with Applications*, 41(7):3497–3505, 2014.

[28] Samuel Danso, Eric Atwell, and Owen Johnson. A comparative study of machine learning methods for verbal autopsy text classification. *arXiv preprint arXiv:1402.4380*, 2014.

[29] Ghulam Mujtaba, Liyana Shuib, Ram Gopal Raj, Retnagowri Rajandram, Khairunisa Shaikh, and Mohammed Ali Al-Garadi. Automatic icd-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PloS one*, 12(2):e0170242, 2017.

[30] Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Peter Serina, David Phillips, Andrea Stewart, Spencer L James, Alireza Vahdatpour, Charles Atkinson, Michael K Freeman, et al. Using verbal autopsy to measure causes of death: the comparative performance of existing methods. *BMC medicine*, 12 (1):5, 2014.

[31] Abraham D Flaxman, Alireza Vahdatpour, Sean Green, Spencer L James, and Christopher JL Murray. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Population health metrics*, 9(1):29, 2011.

[32] Andrew Boulle, Daniel Chandramohan, and Peter Weller. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. *International journal of epidemiology*, 30(3):515–520, 2001.

[33] Maria A Quigley, Daniel Chandramohan, and Laura C Rodrigues. Diagnostic accuracy of physician review, expert algorithms and data-derived algorithms in adult verbal autopsies. *International Journal of Epidemiology*, 28(6):1081–1087, 1999.

[34] Mathew A Mwanyangala, Honorathy M Urassa, Jensen C Rutashobya, Chrisostom C Mahutanga, Angelina M Lutambi, Deodatus V Maliti, Honorati M Masanja, Salim K Abdulla, and Rose N Lema. Verbal autopsy completion rate and factors associated with undetermined cause of death in a rural resource-poor setting of tanzania. *Population health metrics*, 9(1):41, 2011.

[35] Noriah Maraba, Aaron S Karat, Kerrigan McCarthy, Gavin J Churchyard, Salome Charalambous, Kathleen Kahn, Alison D Grant, and Violet Chihota. Verbal autopsy-assigned causes of death among adults being investigated for tb in south africa. *Transactions of the Royal Society of Tropical Medicine and Hygiene*, 110 (9):510–516, 2016.

[36] Gareth James, Daniela Witten, Trevor Hastie, and Robert Tibshirani. *An introduction to statistical learning*, volume 112. Springer, 2013.

[37] Barnaby C Reeves and Maria Quigley. A review of data-derived methods for assigning causes of death from verbal autopsy data. *International journal of epidemiology*, 26(5):1080–1089, 1997.

[38] Juan D Rodriguez, Aritz Perez, and Jose A Lozano. Sensitivity analysis of k-fold cross validation in prediction error estimation. *IEEE transactions on pattern analysis and machine intelligence*, 32(3): 569–575, 2010.

[39] Khaled Fawagreh, Mohamed Medhat Gaber, and Eyad Elyan. Random forests: from early developments to recent advancements. *Systems Science & Control Engineering: An Open Access Journal*, 2(1):602–609, 2014.

[40] Nuntaporn Klinjun, Apiradee Lim, and Kanitta Bundhamcharoen. A logistic regression model for estimating transport accident deaths using verbal autopsy data. *Asia Pacific Journal of Public Health*, 27 (3):286–292, 2015.

[41] Kunal Rajput, Girija Chetty, and Rachel Davey. Phis (protected health information) identification from free text clinical records based on machine learning. In *2017 IEEE Symposium Series on Computational Intelligence (SSCI)*, pages 1–9. IEEE, 2017.

[42] Syed Shariyar Murtaza, Patrycja Kolpak, Ayse Bener, and Prabhat Jha. Automated verbal autopsy classification: using one-against-all ensemble method and naïve bayes classifier. *Gates open research*, 2, 2018.

[43] Riccardo Miotto, Fei Wang, Shuang Wang, Xiaoqian Jiang, and Joel T Dudley. Deep learning for healthcare: review, opportunities and challenges. *Briefings in bioinformatics*, 19(6):1236–1246, 2017.

[44] William H Crown. Potential application of machine learning in health outcomes research and some statistical cautions. *Value in health*, 18(2):137–140, 2015.

[45] John Pestian, Henry Nasrallah, Pawel Matykiewicz, Aurora Bennett, and Antoon Leenaars. Suicide note

classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3: BII–S4706, 2010.

[46] Serena Jeblee, Mireille Gomes, and Graeme Hirst. Multi-task learning for interpretable cause of death classification using key phrase prediction. In *Proceedings of the BioNLP 2018 workshop*, pages 12–17, 2018.

[47] Ekkarat Boonchieng and Khanita Duangchaemkarn. Digital disease detection: Application of machine learning in community health informatics. In *2016 13th International Joint Conference on Computer Science and Software Engineering (JCSSE)*, pages 1–5. IEEE, 2016.

[48] Bevan Koopman, Guido Zuccon, Anthony Nguyen, Anton Bergheim, and Narelle Grayson. Extracting cancer mortality statistics from death certificates: A hybrid machine learning and rule-based approach for common and rare cancers. *Artificial intelligence in medicine*, 2018.

[49] Indrajani Sutedja, Hendro Nindito, Azani Cempaka Sari, et al. Building a mobile health application and heart diagnose to assist patients: Analysis and design. In *Information Management and Technology (ICIMTech), 2017 International Conference on*, pages 19–24. IEEE, 2017.

[50] Khaled Alsabti, Sanjay Ranka, and Vineet Singh. An efficient k-means clustering algorithm. 1997.

[51] Frederick Livingston. Implementation of breiman's random forest machine learning algorithm. *ECE591Q Machine Learning Journal Paper*, 2005.

[52] Jiafan Peng, Shunong Zhang, Dongmu Peng, and Kan Liang. Application of machine learning method in bridge health monitoring. In *Reliability Systems Engineering (ICRSE), 2017 Second International Conference on*, pages 1–7. IEEE, 2017.

[53] Matthew D Byrne. Machine learning in health care. *Journal of PeriAnesthesia Nursing*, 32(5):494–496, 2017.

[54] Bevan Koopman, Sarvnaz Karimi, Anthony Nguyen, Rhydwyn McGuire, David Muscatello, Madonna Kemp, Donna Truran, Ming Zhang, and Sarah Thackway. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC medical informatics and decision making*, 15(1):53, 2015.

[55] Robert E Schapire. Using output codes to boost multiclass learning problems. In *ICML*, volume 97, pages 313–321. Citeseer, 1997.

[56] Samuel J Clark. A guide to comparing the performance of va algorithms. *arXiv preprint arXiv:1802.07807*, 2018.

[57] Samuel J Clark, Zehang Li, and Tyler H McCormick. Quantifying the contributions of training data and algorithm logic to the performance of automated cause-assignment algorithms for verbal autopsy. *arXiv preprint arXiv:1803.07141*, 2018.

[58] Spencer L James, Abraham D Flaxman, and Christopher JL Murray. Performance of the tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Population Health Metrics*, 9(1):31, 2011.

[59] Zehang Richard Li, Tyler H McCormick, and Samuel J Clark. Non-confirming replication of "performance of insilicova for assigning causes of death to verbal autopsies: multisite validation study using clinical diagnostic gold standards," by flaxman et al. *BMC medicine*, 18(1):1–5, 2020.

[60] Nikita Desai, Lukasz Aleksandrowicz, Pierre Miasnikof, Ying Lu, Jordana Leitao, Peter Byass, Stephen Tollman, Paul Mee, Dewan Alam, Suresh Kumar Rathi, 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*, 12(1):20, 2014.

[61] Henry D Kalter, Jamie Perin, and Robert E Black. Validating hierarchical verbal autopsy expert algorithms in a large data set with known causes of death. *Journal of global health*, 6(1), 2016.

[62] Daniel Chandramohan, Philip Setel, and Maria Quigley. Effect of misclassification of causes of death in verbal autopsy: can it be adjusted? *International Journal of Epidemiology*, 30(3):509–514, 2001.

[63] Samuel Danso, Eric Atwell, and Owen Johnson. Linguistic and statistically derived features for cause of death prediction from verbal autopsy text. In *Language processing and knowledge in the web*, pages 47–60. Springer, 2013.

[64] SA Danso, Owen Johnson, A Ten Asbroek, S Soromekun, K Edmond, C Hurt, L Hurt, C Zandoh, C Tawiah, J Fenty, et al. A semantically annotated verbal autopsy corpus for automatic analysis of cause of death. *ICAME Journal*, 37, 2013.

[65] Samuel Odei Danso. *Text Analytics to Predict Time and Cause of Death from Verbal Autopsies*. PhD thesis, University of Leeds, 2015.

[66] Tom Young, Devamanyu Hazarika, Soujanya Poria, and Erik Cambria. Recent trends in deep learning based natural language processing. *arXiv preprint arXiv:1708.02709*, 2017.

[67] Ronan Collobert and Jason Weston. A unified architecture for natural language processing: Deep neural networks with multitask learning. In *Proceedings of the 25th international conference on Machine learning*, pages 160–167. ACM, 2008.

[68] Carol Friedman, Lyudmila Shagina, Yves Lussier, and George Hripcsak. Automated encoding of clinical documents based on natural language processing. *Journal of the American Medical Informatics Association*, 11(5):392–402, 2004.

[69] Vendhan Gajalakshmi and Richard Peto. Suicide rates in rural tamil nadu, south india: verbal autopsy of 39 000 deaths in 1997–98. *International journal of epidemiology*, 36(1):203–207, 2007.

[70] Cannannore Nidhi Kamath, Syed Saqib Bukhari, and Andreas Dengel. Comparative study between traditional

machine learning and deep learning approaches for text classification. In *Proceedings of the ACM Symposium on Document Engineering 2018*, pages 1–11, 2018.

[71] Wenfeng Tian, Jun Li, and Hongguang Li. A method of feature selection based on word2vec in text categorization. In *2018 37th Chinese Control Conference (CCC)*, pages 9452–9455. IEEE, 2018.

[72] Long Ma and Yanqing Zhang. Using word2vec to process big text data. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 2895–2897. IEEE, 2015.

[73] Marwa Naili, Anja Habacha Chaibi, and Henda Hajjami Ben Ghezala. Comparative study of word embedding methods in topic segmentation. *Procedia Computer Science*, 112:340–349, 2017.

[74] Eissa M Alshari, Azreen Azman, Shyamala Doraisamy, Norwati Mustapha, and Mustafa Alkeshr. Improvement of sentiment analysis based on clustering of word2vec features. In *2017 28th International Workshop on Database and Expert Systems Applications (DEXA)*, pages 123–126. IEEE, 2017.

[75] Nabil Alami, Mohammed Meknassi, and Noureddine En-nahnahi. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. *Expert Systems with Applications*, 2019.

[76] Alexander Kirillov, Dmitrij Schlesinger, Walter Forkel, Anatoly Zelenin, Shuai Zheng, Philip Torr, and Carsten Rother. Efficient likelihood learning of a generic cnn-crf model for semantic segmentation. *arXiv preprint arXiv:1511.05067*, 2015.

[77] Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE international conference on computer vision*, pages 1–9, 2015.

[78] Yoon Kim. Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*, 2014.

[79] Christopher JL Murray, Rafael Lozano, Abraham D Flaxman, Alireza Vahdatpour, and Alan D Lopez. Robust metrics for assessing the performance of different verbal autopsy cause assignment methods in validation studies. *Population health metrics*, 9(1):28, 2011.

[80] Gaoxia Jiang and Wenjian Wang. Error estimation based on variance analysis of k-fold cross-validation. *Pattern Recognition*, 69:94–106, 2017.

[81] Tzu-Tsung Wong. Performance evaluation of classification algorithms by k-fold and leave-one-out cross validation. *Pattern Recognition*, 48(9):2839–2846, 2015.

[82] Samuel Danso, Eric Atwell, Owen Johnson, Guus ten Asbroek, Karen Edmond, C Hurt, L Hurt, C Zandoh, C Tawiah, Z Hill, et al. A verbal autopsy corpus for machine learning of cause of death. In *Proceedings of the Corpus Linguistics Conference*, 2011.

[83] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1631–1642, 2013.

[84] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[85] Samantha Herrera, Yeetey Enuameh, George Adjei, Kenneth Ayuurebobi, Kwaku Poku Asante, Osman Sankoh, Seth Owusu-Agyei, Yazoume Yé, et al. A systematic review and synthesis of the strengths and limitations of measuring malaria mortality through verbal autopsy. *Malaria journal*, 16(1):421, 2017.

VIII. Figure Legends

This section presents a list of figures in this paper.

- Fig. 1 ANN structure of single neuron.
- Fig. 2 Comparison of ANN and deep learning architecture (Young 2017).
- Fig. 3 Convolutional neural network framework (Young 2017).
- Fig. 4 A confusion matrix with correct predictions, True Positives (TP) and True Negatives (TN), and incorrect predictions, False Positives (FP) and False Negatives (FN).
- Fig. 5 Recurrent neural network network framework (Young 2017).
- Fig. 6 Long term short memory framework (Young 2017).
- Fig. 7 Distribution of initial searched journal articles vs conference papers.
- Fig. 8 Process flowchart of the systematic literature survey methodology.
- Fig. 9 Distribution of VA primary studies that include pre-processing.
- Fig. 10 Distribution of VA primary studies that include transparency.
- Fig. 11 Distribution of VA related studies that include data balancing.
- Fig. 12 Distribution of feature extraction techniques reviewed. LEGEND: BAG OF WORDS (BOW), N-GRAMS (NG), HYBRID (HB), WORD2VEC (WV), CONCEPT BASED (CB), GRAPH BASED (GB), RULE BASE (RB), DOC2VEC (DV), EXPERT DRIVEN (ED), SENTENCE2VEC (SV), GLOVE (GV).
- Fig. 13 Distribution of feature value representation schemes. LEGEND: BINARY REPRESENTATION (BR), TERM FREQUENCY (TF), DOCUMENT FREQUENCY (DF), TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF), NORMALISED TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (NTF-IDF).
- Fig. 14 Distribution of feature selection techniques. LEGEND: PEARSON CORRELATION (PC), CHI-SQUARED (CS), INFORMATION GAIN (IG), PRINCIPAL COMPONENT ANALYSIS (PCA),FISHER MARKOV SELECTOR (FMS), IMPROVED GLOBAL FEATURE SELECTION (IGFS), LOCAL SEMI SUPERVISED (LSS), NATURAL LANGUAGE PROCESSING (NLP).
- Fig. 15 Distribution of VA related studies in comparison to other fields used in this review.
- Fig. 16 Distribution of journal articles and conference articles used in this review.
- Fig. 17 Distribution of evaluation metrics from primary studies reviewed. LEGEND: SPECIFICITY (SP), PRECISION (PR), RECALL (RE), F-SCORE (FS), ACCURACY (AC), CSFM (CS), CCC, PCCC, KAPPA (KA), CASE SIMILARITY MEASURE (CSM), TOP CAUSE (TC), TOP3COD (T3), AREA UNDER CURVE (AUC).
- Fig. 18 Tag cloud of this systematic review.
- Fig. 19 Distribution of statistical and ML models used in primary studies. LEGEND: VA ALGORITHMS (VA), LOGISTIC REGRESSION (LR), RANDOM FORESTS (RF), DECISION TREE (DT), K- NEAREST NEIGHBOR (K-NN), SUPPORT VECTOR MACHINE (SVM), ARTIFICIAL NEURAL NETWORK (ANN), NAIVE BAYES CLASSIFIER (NBC), K-MEANS CLUSTERING (K-MEANS), RULE BASED (RB), ENSEMBLE CLASSIFIER (EC), CONVOLUTIONAL NEURAL NETWORK (CNN), GATED RECURRENT NEURAL NETWORK (GRN).
- Fig. 20 The guiding machine learning VA that is proposed in this literature survey.

Figures

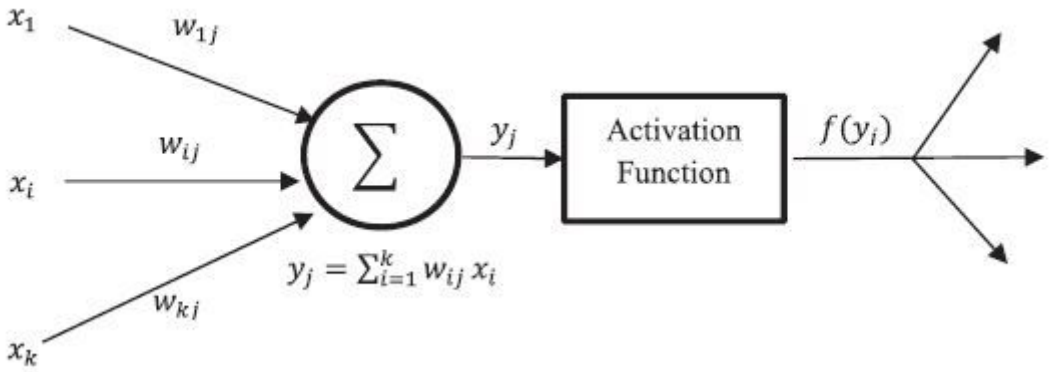

Figure 1

ANN structure of single neuron.

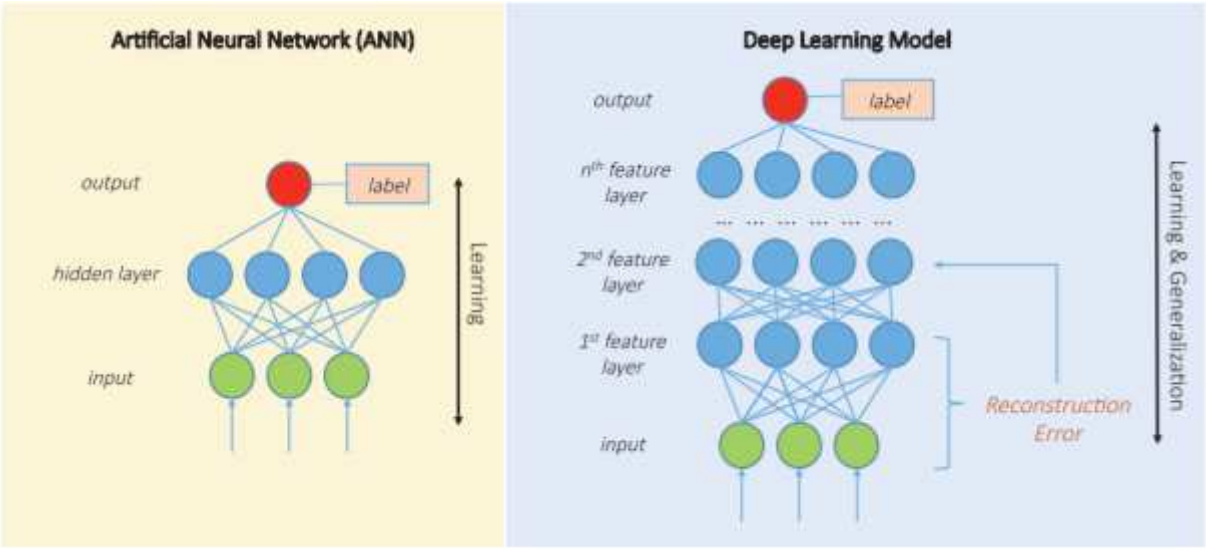

Figure 2

ComparisonofANNanddeeplearningarchitecture (Young 2017).

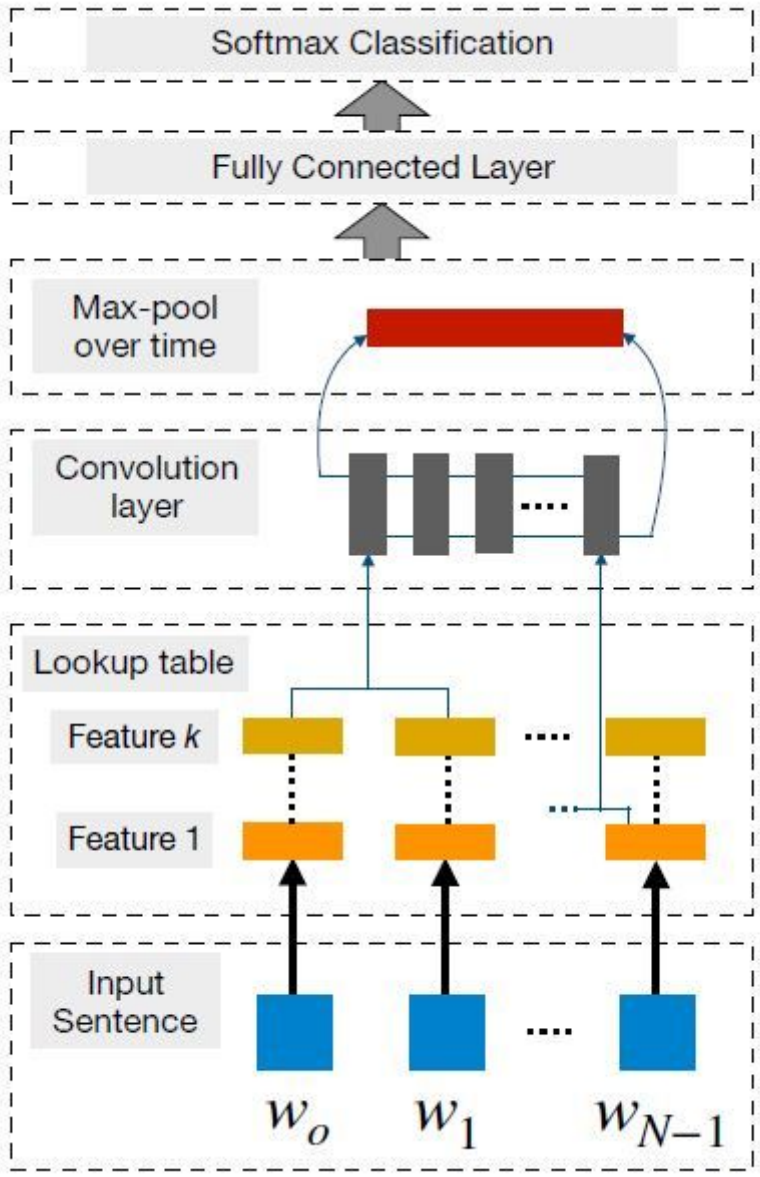

Figure 3

Convolutional neural network framework (Young 2017).

Actual labels	Predicted labels		
		Positives	Negatives
	Positives	TP	FP
	Negatives	FN	TN

Figure 4

A confusion matrix with correct predictions, True Positives (TP) and True Negatives (TN), and incorrect predictions, False Positives (FP) and False Negatives (FN).

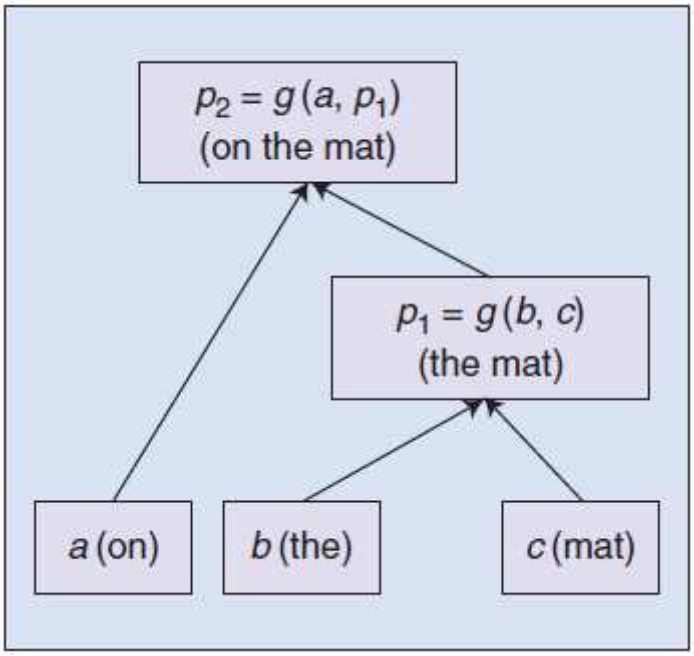

Figure 5

Recurrent neural network network framework (Young 2017).

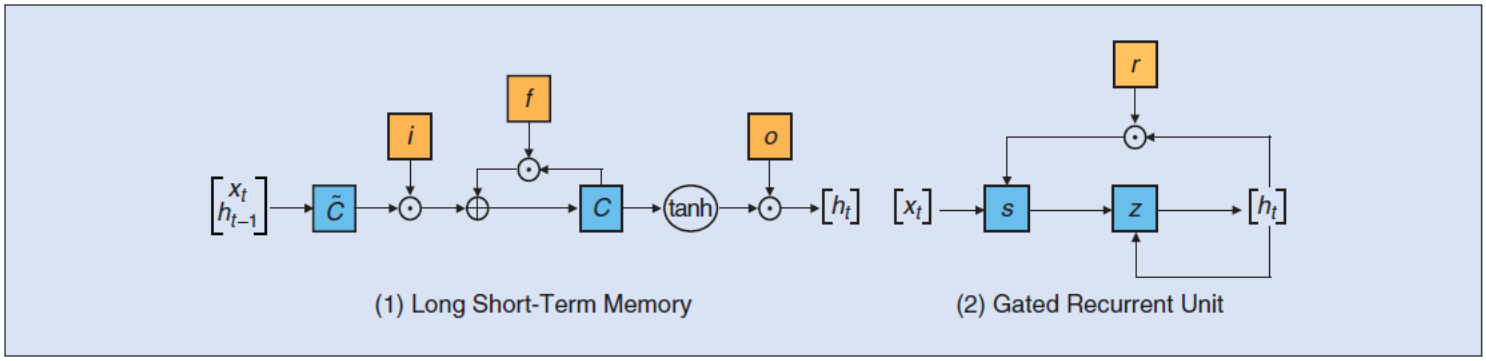

Figure 6

Long term short memory framework (Young 2017).

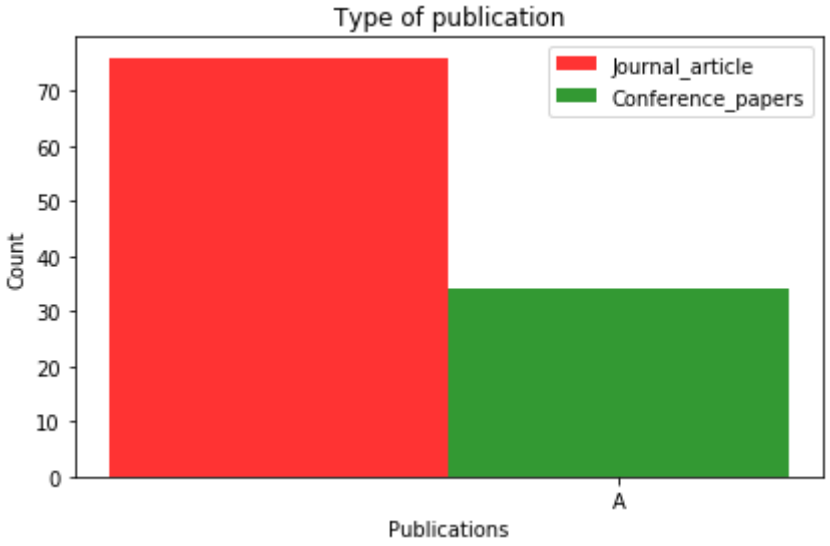

Figure 7

Distribution of initial searched journal articles vs conference papers.

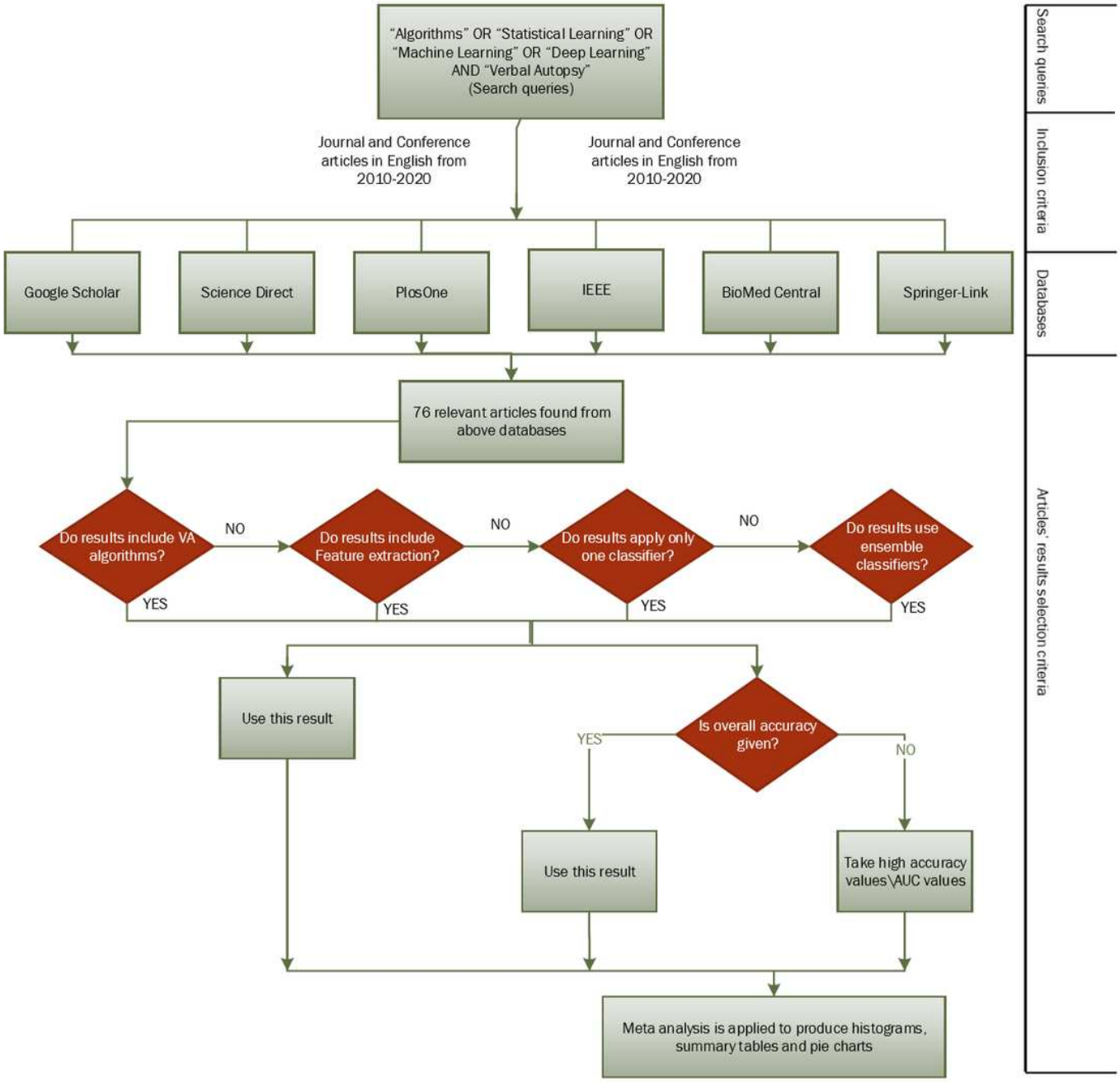

Figure 8

Process flowchart of the systematic literature survey methodology.

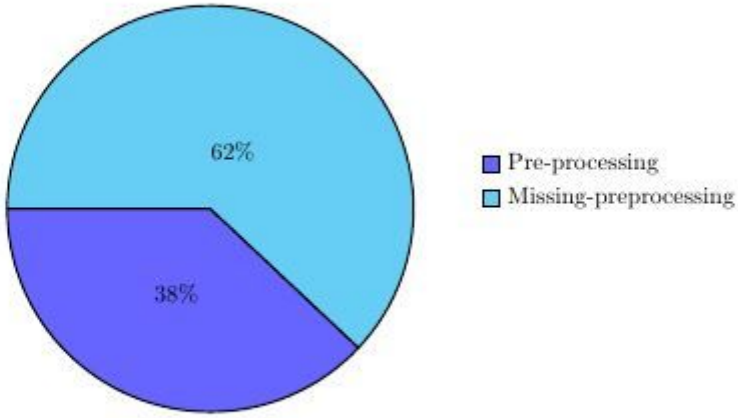

Figure 9

Distribution of VA primary studies that include pre-processing.

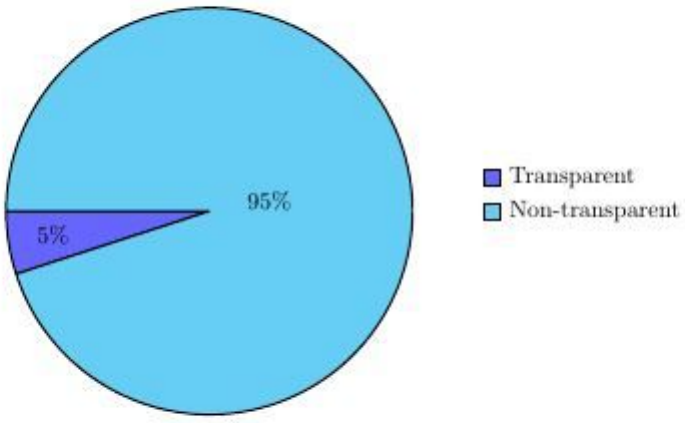

Figure 10

Distribution of VA primary studies that include transparency.

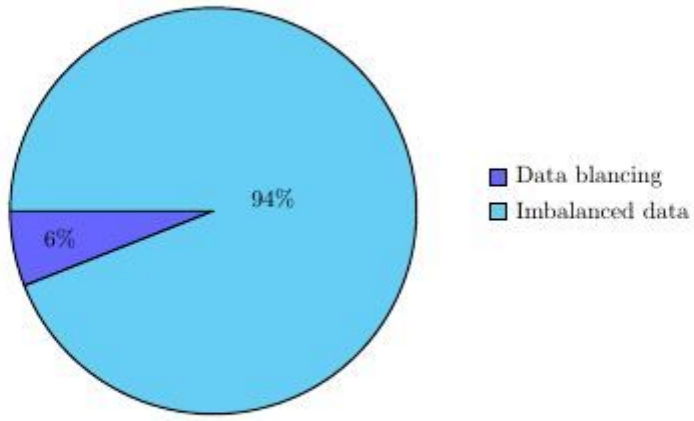

Figure 11

Distribution of VA related studies that include data balancing.

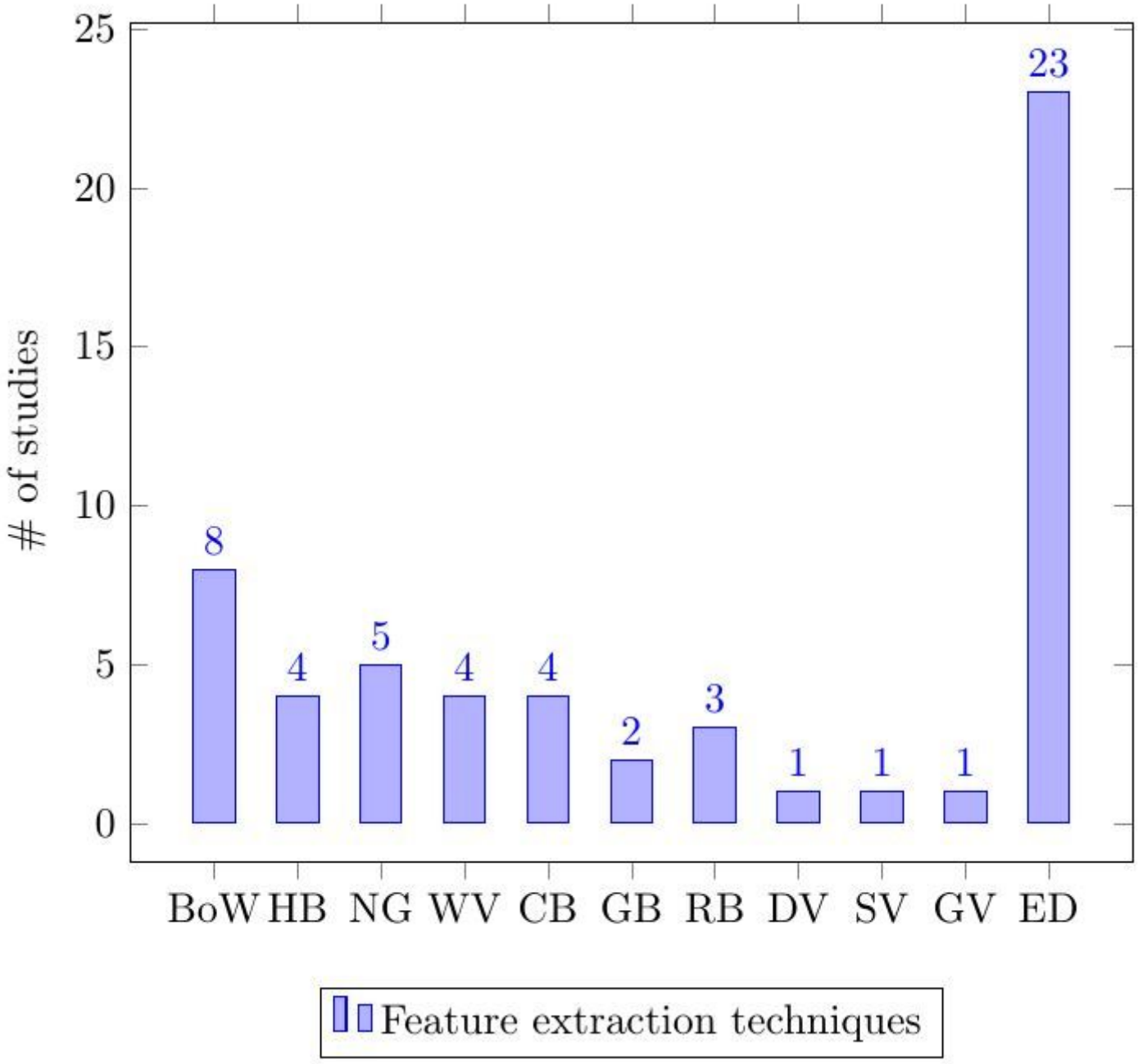

Figure 12

Distribution of feature extraction techniques reviewed. LEGEND: BAG OF WORDS (BOW), NGRAMS (NG), HYBRID (HB), WORD2VEC (WV), CONCEPT BASED (CB), GRAPH BASED (GB), RULE

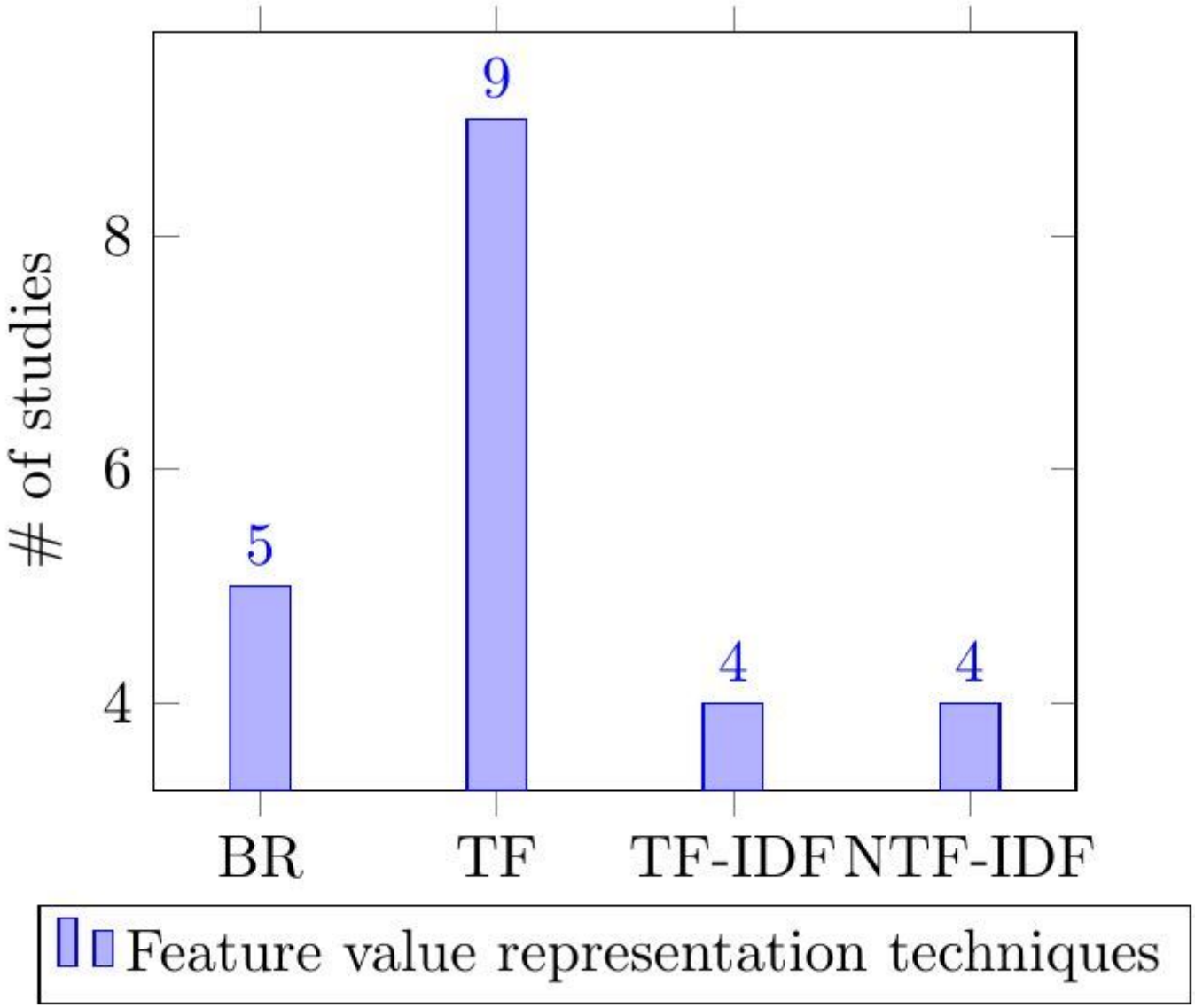

Figure 13

Distribution of feature value representation schemes. LEGEND: BINARY REPRESENTATION (BR), TERM FREQUENCY (TF), DOCUMENT FREQUENCY (DF), TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (TF-IDF), NORMALISED TERM FREQUENCY-INVERSE DOCUMENT FREQUENCY (NTF-IDF).

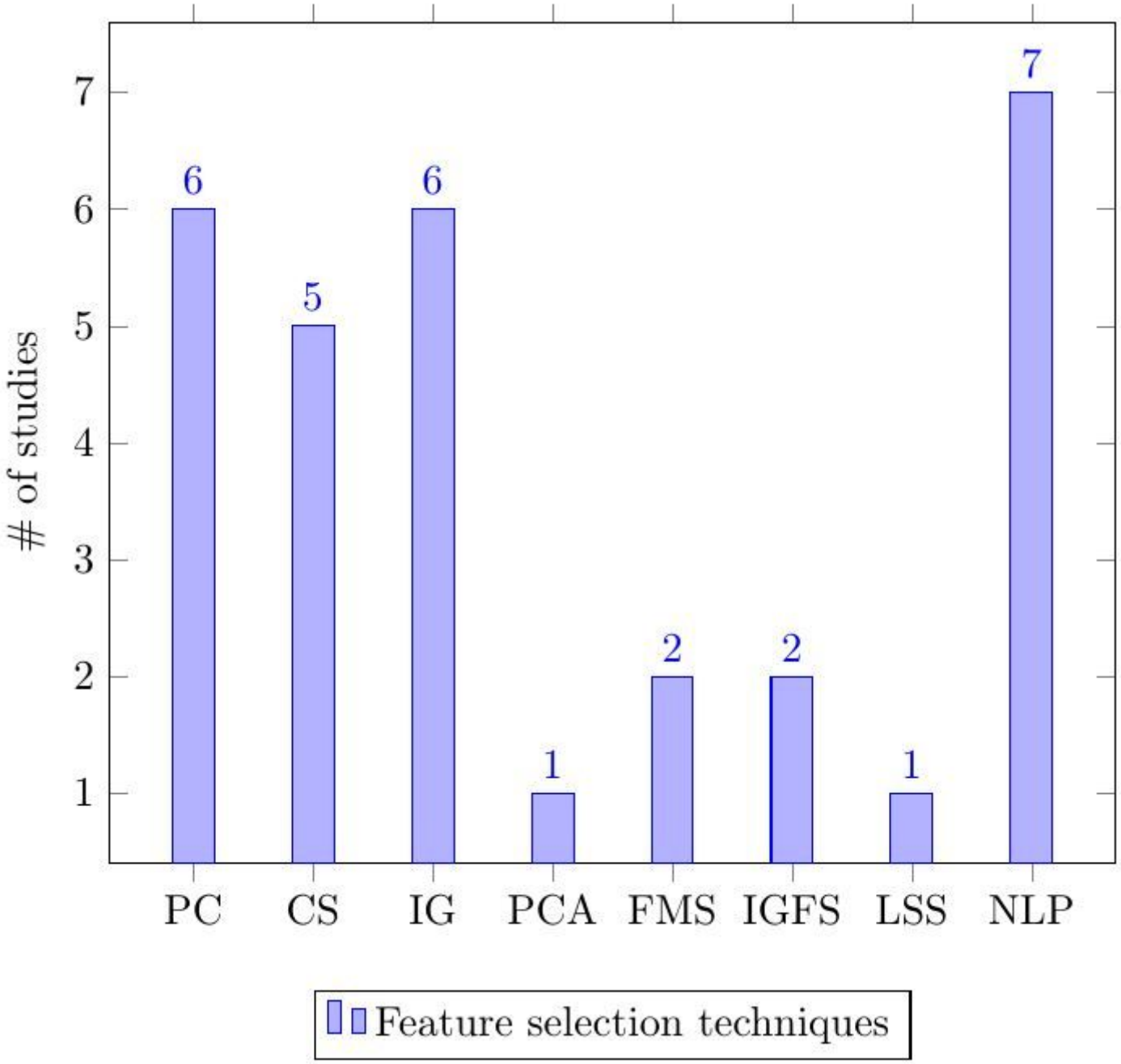

Figure 14

Distribution of feature selection techniques. LEGEND: PEARSON CORRELATION (PC), CHISQUARED (CS), INFORMATION GAIN (IG), PRINCIPAL COMPONENT ANALYSIS (PCA),FISHER MARKOV SELECTOR (FMS), IMPROVED GLOBAL FEATURE SELECTION (IGFS), LOCAL SEMI SUPERVISED (LSS), NATURAL LANGUAGE BASE (RB), DOC2VEC (DV), EXPERT DRIVEN (ED), SENTENCE2VEC (SV), GLOVE (GV).

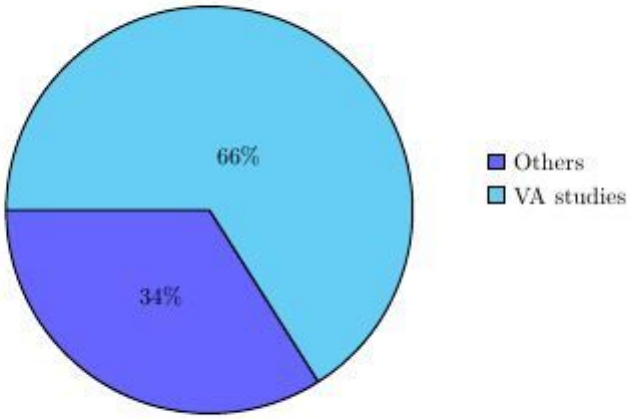

Figure 15

Distribution of VA related studies in comparison to other fields used in this review.

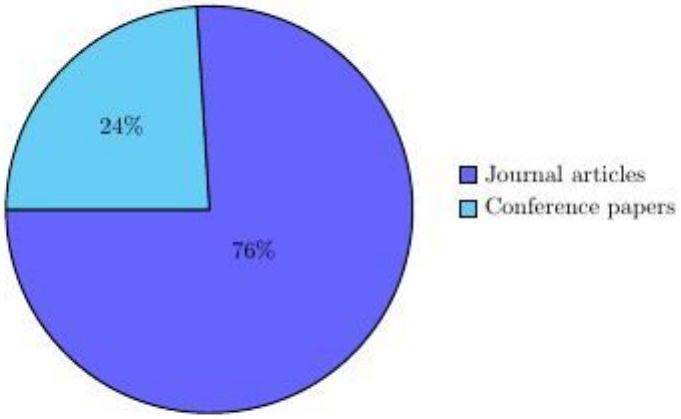

Figure 16

Distribution of journal articles and conference articles used in this review.

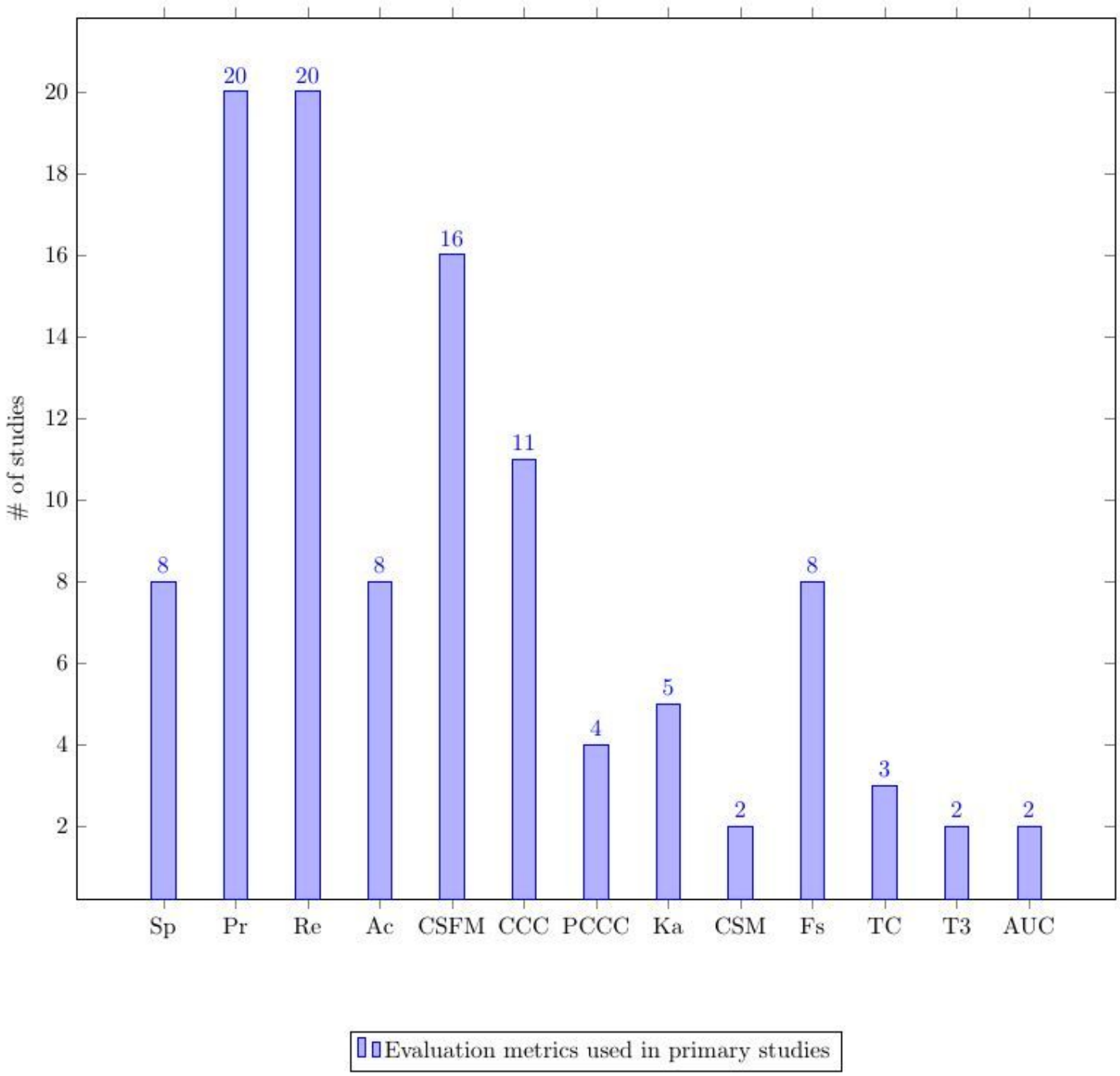

Figure 17

Distribution of evaluation metrics from primary studies reviewed. LEGEND: SPECIFICITY (SP), PRECISION (PR), RECALL (RE), F-SCORE (FS), ACCURACY (AC), CSFM (CS), CCC, PCCC, KAPPA (KA), CASE SIMILARITY MEASURE (CSM), TOP CAUSE (TC), TOP3COD (T3), AREA UNDER CURVE (AUC).

Figure 18

Tag cloud of this systematic review.

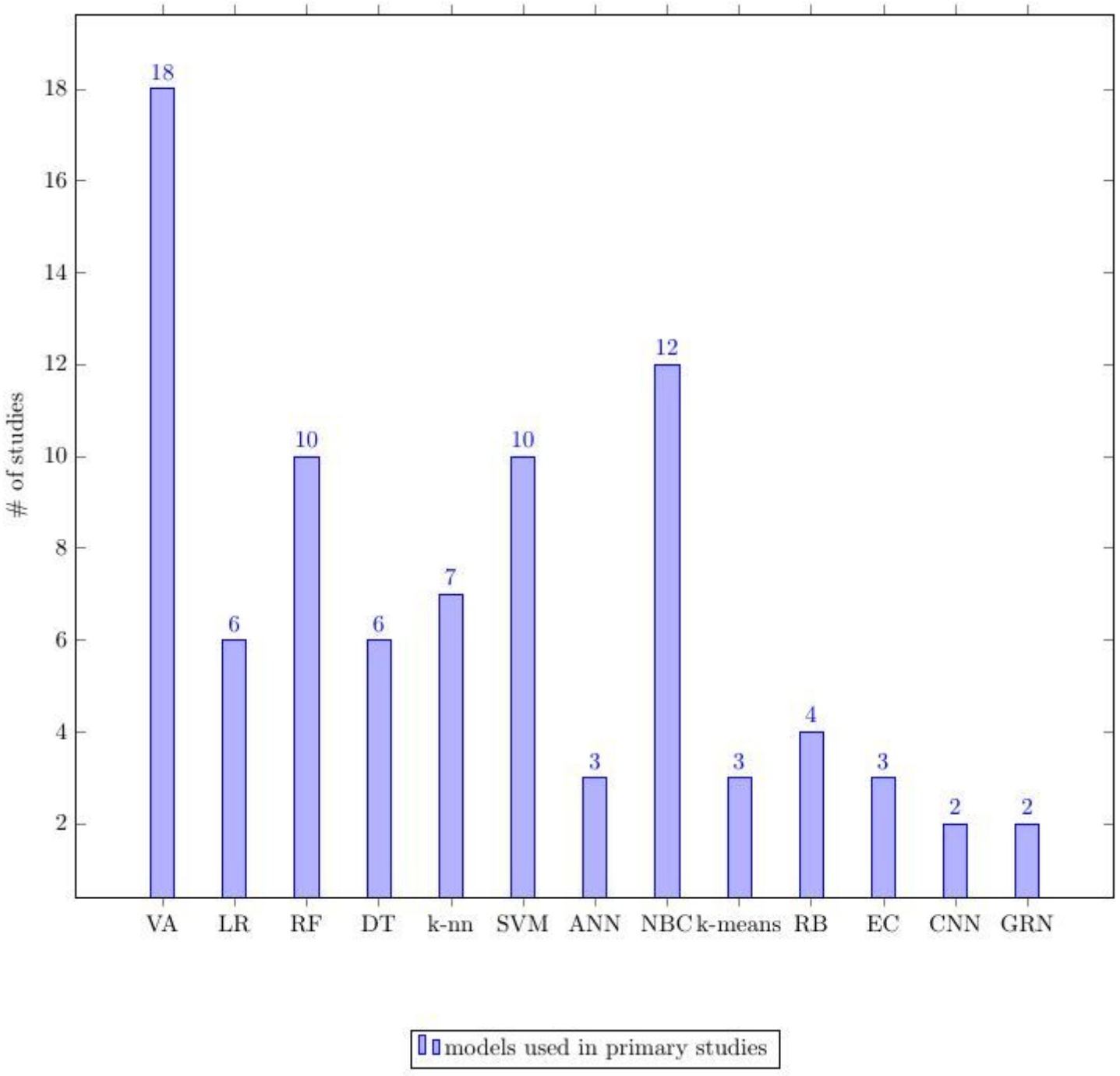

Figure 19

Distribution of statistical and ML models used in primary studies. LEGEND: VA ALGORITHMS (VA), LOGISTIC REGRESSION (LR), RANDOM FORESTS (RF), DECISION TREE (DT), K- NEAREST NEIGHBOR (K-NN), SUPPORT VECTOR MACHINE (SVM), ARTIFICIAL NEURAL NETWORK (ANN), NAIVE BAYES CLASSIFIER (NBC), K-MEANS CLUSTERING (K-MEANS), RULE BASED (RB), ENSEMBLE CLASSIFIER (EC), CONVOLUTIONAL NEURAL NETWORK (CNN), GATED RECURRENT NEURAL NETWORK (GRN).

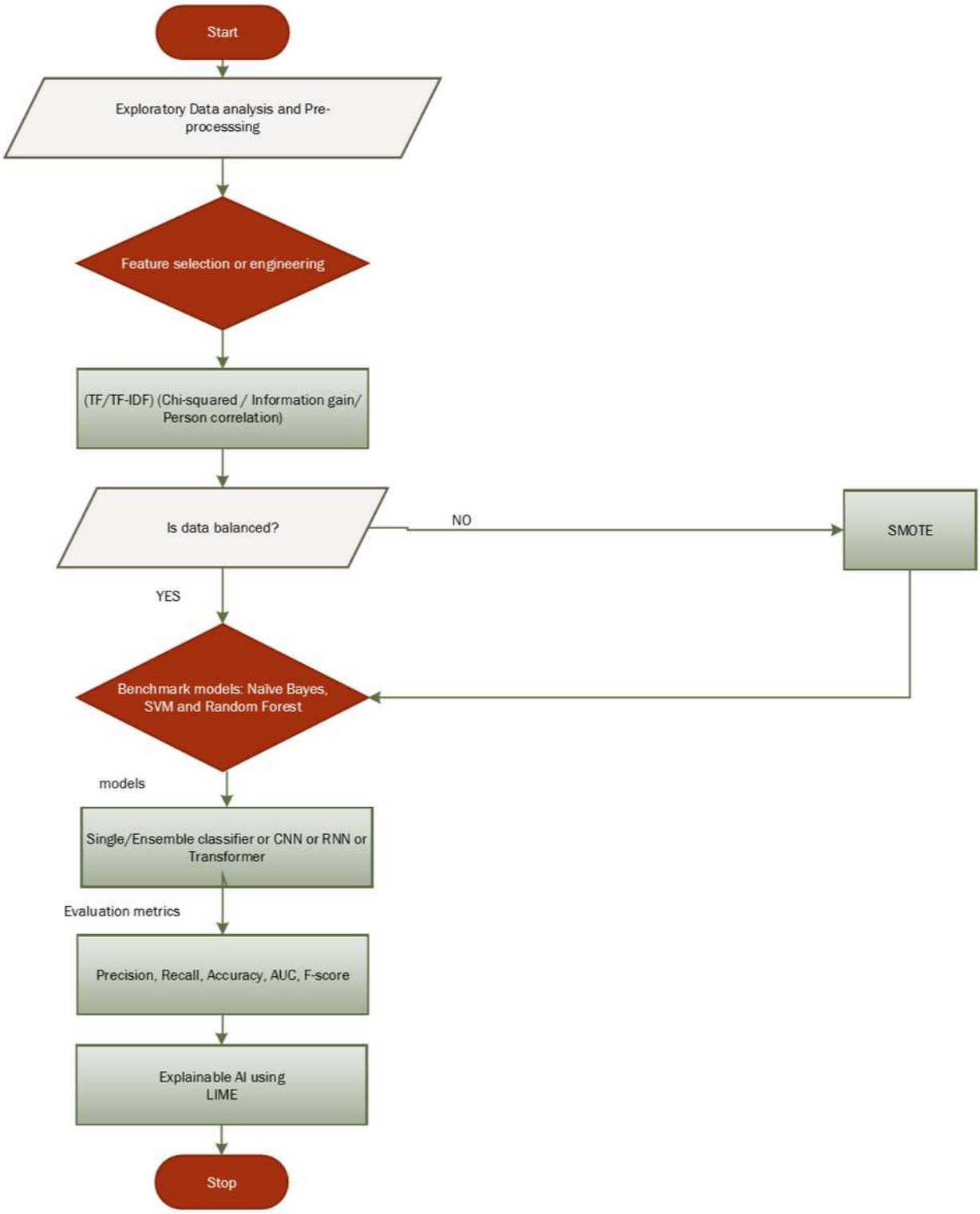

Figure 20

The guiding machine learning VA that is proposed in this literature survey.