

ORIGINAL RESEARCH

Open access

Interpretable Deep Learning for Verbal Autopsy Analysis: A Stacked Ensemble Approach to Cause-of-Death Determination

Amira Khalil^{1*}, Sara Fathy¹, Mahmoud Adel², Hany Nabil³, Rania Mostafa⁴

Abstract

Verbal autopsies (VA) serve as a frequent mechanism within Low- and Medium-Income Countries (LMIC) to ascertain the cause of death (CoD) for individuals who pass away outside of clinical environments, relative to the established international gold standard of physician medical certification. Field interviewers gather detailed testimonies from the families of the deceased concerning the events and circumstances preceding the fatality. This gathered information is documented in text format, referred to as VA narratives. These textual records contain granular details suitable for identifying the CoD. Nonetheless, this methodology continues to be a manual operation characterized by high costs, inconsistency, prolonged processing times, and subjectivity (making it susceptible to errors), among other limitations. Consequently, these disadvantages undermine the VA reporting infrastructure, which is otherwise crucial for defining public health agendas and populating civil registration frameworks.

To address this operational gap, the present study introduces interpretable deep learning (DL) methodologies designed to analyze VA narratives and deliver CoD predictions in an efficient, transparent, economical, and precise manner. The verification of our DL models is conducted via optimization and performance accuracy machine learning (ML) curves mapped against the volume of training samples. Our models achieved training set accuracies of (LSTM = 76.11%, CNN = 76.35%, and SEDL = 82.1%), validation accuracies of (LSTM = 67.05%, CNN = 66.16%, and SEDL = 82%), and test set accuracies of (LSTM = 67%, CNN = 66.2%, and SEDL = 82%). Additionally, Local Interpretable Model-agnostic Explanations (LIME) are utilized to enhance the interpretability of the outcomes, thereby fostering confidence in automated healthcare solutions.

This work offers robust deep learning strategies for CoD determination via VAs, where the stacked ensemble deep learning (SEDL) model achieved optimal performance, outperforming both Long Short-Term Memory (LSTM) and Convolutional Neural Networks (CNN). Our empirical evidence suggests that ensemble DL paradigms can be successfully integrated into the CoD classification workflow to assist medical experts in reaching a diagnosis. Ultimately, this integration aims to compress the diagnostic turnaround time for physicians reviewing the narratives, while simultaneously reducing operational expenses and mitigating human error. The scope of this study was constrained by the volume of training data required for model optimization and the substantial lexical diversity present within the unstructured text.

Keywords Machine learning, LIME, Deep learning, CNN, Cause of death, LSTM

*Correspondence:

Amira Khalil
amira.khalil@alexu.edu.eg

¹ Department of Verbal Autopsy and Interpretable Deep Learning, Faculty of Medicine, Alexandria University, Alexandria, Egypt

² Department of Cause-of-Death Determination and Stacked Ensembles, Faculty of Medicine, Cairo University, Cairo, Egypt

³ Department of Interpretable Models for Public Health, Faculty of Medicine, Mansoura University, Mansoura, Egypt

⁴ Department of Ensemble Learning and XAI, Faculty of Medicine, Helwan University, Cairo, Egypt

Introduction

Upward of 65 percent of the global population does not have access to high-quality data regarding cause of death (CoD), which represents two thirds of the sixty million annual global fatalities lacking a formally certified medical cause [1-3]. Such statistical insights remain vital for maintaining civil registration databases, drafting policies, and reinforcing national health strategies. The majority of these unrecorded deaths take place within Low- and Medium-Income Countries (LMIC) due to their occurrence outside formal medical centers [4]. Consequently, the verbal autopsy (VA) instrument is deployed to ascertain CoD in these scenarios.

The VA paradigm involves non-clinical personnel conducting structured interviews with the primary relatives of the deceased to extract specific data regarding the circumstances and chronological events leading up to the CoD. The compiled qualitative data, termed VA narratives, are digitally recorded as text and evaluated by two independent medical practitioners to deduce the likely CoD; in instances of disagreement, a third physician adjudicates the case, an operational framework designated as physician-coded verbal autopsy (PCVA) [3]. A comprehensive breakdown of the traditional VA workflow is detailed in [5]. Although PCVA represents the primary reference standard for CoD evaluation, the protocol is financially burdensome, slow, highly variable, and subjective (vulnerable to inconsistencies during data elicitation), among numerous other limitations.

This has spurred the development of alternative algorithmic approaches that leverage probabilistic matrices, tariff scoring systems, and basic statistical computations to classify the CoD [6, 7]. These methodologies are collectively classified as computer-coded verbal autopsy (CCVA). Contemporary literature has predominantly concentrated on statistical frameworks evaluating CoD categorization through CCVA configurations applied to VA narratives [7-11]. Nonetheless, these systems often yield suboptimal performance due to their inability to effectively handle the unstructured text narratives provided as classification inputs. As a result, standard CCVA methods fall short of guiding healthcare planning effectively, given their failure to produce robust diagnostic evidence in environments with scarce expert verification [1].

Conversely, parallel research efforts have explored machine learning (ML) models, which have yielded superior performance indicators compared to traditional CCVA methods. Significant investigation has centered on employing various shallow ML algorithms trained on highly high-dimensional spaces (characterized by extensive sets of predictors and target outcomes) containing sparse features (where variables are not distributed across uniform scales) to extract actionable intelligence from VA text records [3, 12-20]. A large portion of these ML methods exhibit performance deficits owing to the curse of high dimensionality induced by absent feature scaling, suboptimal feature engineering, inadequate text preprocessing pipelines, and unaddressed class imbalances.

Technological advancements, particularly across the domains of ML and natural language processing (NLP), have unlocked sophisticated capabilities and improved efficiencies for interpreting and categorizing unstructured text narratives. In the current paper, we introduce robust deep learning (DL) architectures driven by NLP that are customized to parse VA text files efficiently and project the most probable CoD. Additionally, this study intends to gauge the systemic impact of these DL frameworks within broader healthcare contexts and mortality statistics surveillance. Through this approach, we seek to elevate CoD classification accuracy by exploiting the advanced hierarchical features of ensemble DL paradigms against the baseline configurations detailed in [21].

Our proposed DL models also offer explicit insights regarding the underlying decision paths for CoD predictions, an essential requirement for establishing transparency and user trust for artificial intelligence diagnostics applied within public health domains. Consequently, this facilitates an exploration of real-world SEDL deployment, paving the way for testing its generalization across external contexts—such as distinct demographics, geographical territories, and alternative cohorts—thereby uncovering novel domain insights and mapping the underlying semantic associations connecting raw VA narratives to their respective CoD categories.

While DL approaches demonstrate superior predictive accuracy, they remain bounded by architectural complexity and constraints regarding model interpretability (the capacity to trace the causal mechanisms behind a specific model output). Notwithstanding these hurdles, DL models deliver highly promising outputs by virtue of computational efficiency, outstanding performance metrics, automated feature extraction, and the capacity to facilitate end-to-end learning directly from complex, multi-modal data streams [22]. By contrast, traditional statistical tools and classical ML pipelines require manual feature engineering to generate representative data attributes prior to the training of a predictive model. To date, academic literature remains sparse concerning the application of these advanced ML architectures, specifically DL frameworks, to address challenges inherent in verbal autopsy analysis.

This investigation furthers the transparency, reliability, and accountability of AI-enabled solutions within clinical and public health frameworks. Incorporating an explainable methodology of this nature into CoD classification promises to minimize diagnostic errors, mitigate algorithmic bias, accelerate system turnaround time, and lower logistical expenses, among other advantages. Consequently, it establishes a framework for productive human-AI collaboration during data interpretation and clinical annotation, ultimately uplifting the overall data quality of VA registries and strategic decision-making crucial for supporting civil records and defining public health priorities.

Deep Learning in VA

Advanced DL functionalities, including automated feature extraction and architectural engineering, have catalyzed the development of automated diagnostic frameworks—an evolution that was largely unattainable using traditional ML approaches [23]. Computational DL models have yielded promising outcomes and have been widely integrated into the medical domain, as demonstrated by the investigations of [24–27]. Additional research by [27–29] similarly highlights the efficacy of DL methodologies. Furthermore, Ref. [30] noted that DL models can enhance classical ML benchmarks across numerous disciplines, including speech recognition, pharmaceutical discovery, and computer vision.

Moreover, DL frameworks configured with optimized hidden layers have been engineered to uncover latent patterns that remain obscured when evaluated by conventional statistical and standard ML tools. These architectural frameworks facilitate the learning of raw data representations across multiple hierarchical layers of abstraction, even when executing operations with relatively few processing layers [31]. The feature engineering paradigm within DL—encompassing feature extraction, representation, and selection—is executed via a unified end-to-end framework that learns directly and autonomously from the training observations. As a consequence, these sophisticated models hold strong relevance for processing VA registries characterized by high-dimensional, sparse data matrices.

Historically, the Convolutional Neural Network (CNN) paradigm was designed for image recognition tasks leveraging multi-dimensional data grids. Nonetheless, subsequent exploratory studies with this architecture revealed superior metrics when adapted for text classification tasks using one-dimensional arrays and character-level inputs [29, 32]. Multiple investigators have deployed the standard Recurrent Neural Network (RNN), with certain researchers modifying the framework for sentiment analysis to achieve notable performance [33–38]. Other studies, such as those by [30, 39], integrated a hybrid CNN and RNN configuration, yielding strong outcomes within a hierarchical implementation.

Ref. [40] evaluated a multi-task learning architecture for CoD categorization. Their design combined a CNN and Linear Discriminant Analysis (LDA) for topic segmentation. The resulting predicted keyphrase clusters surpassed individual LDA and CNN formulations in extracting diagnostic terms from the underlying VA records. Specifically, they utilized a feed-forward CNN network that ingested word embedding vectors, optimizing the system via 10-fold cross-validation to produce distinct LDA topical categories. Their configuration achieved a precision of 0.779, a recall of 0.778, and an F1-score of 0.774. Nonetheless, their analysis was constrained by a small sample size and featured word clusters that exhibited disproportionate frequencies and lengths, thereby introducing unrepresentative data features.

Ref. [41] investigated the utility of character embeddings on VA records to evaluate whether such an approach could elevate CoD classification metrics. Utilizing four separate datasets spanning distinct clinical categories, they determined that character-level features boost classification performance when applied to smaller cohorts, as the underlying networks can successfully manage out-of-vocabulary terms and spelling variations. They observed a meaningful performance increase when deploying models such as a CNN. However, their experimentation was restricted to limited datasets containing fewer than 1000 total cases, which also fluctuated in sample size.

Ref. [42] utilized explainable DL frameworks to deduce CoD from VA text by benchmarking logistic regression, a naive Bayes classifier, a random forest ensemble, and two distinct configurations of RNN DL models. Their empirical trials spanned three datasets; however, competitive performance metrics were isolated to the neonatal cohort evaluated by the DL model, which yielded a precision of 63.2%, an accuracy of 63.0%, a recall of 63.0%, and an F-score of 61.3%.

Ref. [43] experimented with the prediction of mortality trajectories during the clinical treatment of SARS-COV-2 patients using a mix of DL and ML approaches. They reported an accuracy of 97.15% for the DL architectures, compared to 92.15% for random forest, 93.4% for k-nearest neighbor, and 99.7% for the XGBoost ensemble classifier. Ref. [44] conducted an evaluation of text narratives concerning railway accidents by pairing DL-based distributed vector representations with an RNN and a CNN. Their configuration achieved a peak accuracy of 75% and an F-score of 0.65 by combining word2vec embeddings with a CNN classifier.

Ensemble DL

Beyond the standard individual DL configurations reviewed above, several investigations have leveraged ensemble DL frameworks (which integrate multiple separate neural models) to achieve superior generalization performance on unseen data. The core mechanism of ensemble DL relies on an averaging or majority voting protocol across the aggregated architectures to formulate a final model prediction.

Ref. [23] performed a study applying SEDL with a CNN backend to diagnose pediatric pneumonia via chest X-ray imagery. Their methodology merged conventional ML tools with advanced DL pipelines. They reported an accuracy of 98.3%, a precision of 99.29%, a recall of 98.36%, an F1-score of 98.83%, and an AUC of 98.24%. Based on these metrics, they concluded that their stacked ensemble configuration achieved optimal operational efficiency, suggesting its utility as a diagnostic aid for medical practitioners.

A comparable study detailed in [45] reinforced the structural robustness of ensemble architectures in achieving high-tier performance metrics. Nonetheless, they emphasized the critical necessity of mitigating class imbalances to ensure reliable outcomes. Ref. [46] implemented an ensemble DL system focused on detecting COVID-19 cases from chest radiographs. They recorded an overall accuracy of 95% and affirmed that ensemble systems can generate superior diagnostic capabilities. However, they noted the ongoing requirement for larger training sample sizes to optimize model training.

Ref. [47] completed a systematic review covering various taxonomies of ensemble DL techniques and verified that ensemble frameworks consistently deliver superior predictive results. Even so, they argued that these large DL structures are difficult to train effectively when sample sizes are small, are inherently complex, and present notable challenges during base model selection. The authors further highlighted the need for prospective research to define systematic strategies for establishing the ideal number of base learners, formulate objective criteria for model selection, and evaluate diverse feature-fusion protocols.

Ref. [23] argued that stacking represents an exceptionally powerful ensemble DL strategy, which functions by assembling a tier of predictions generated by independent base classifiers. These intermediate predictions from the primary models are then treated as higher-level features, which serve as the direct inputs for training a meta-learner responsible for

outputting the final classification decision. This approach generates substantial performance gains because stacking successfully aggregates the complementary strengths of the individual base classifiers.

Lastly, Ref. [48] reported on the implementation of ensemble DL architectures tasked with classifying cardiovascular diseases utilizing patient records from Mexico. They recorded accuracy rates and F1-scores resting tightly within the 91% to 96% range. From these findings, they concluded that ensemble DL paradigms can sustain excellent performance thresholds, making them highly viable for real-time diagnostic tracking and clinical reporting within medical environments.

Explainable AI

While DL approaches demonstrate exceptional promise across diverse applications, their architectural complexity severely limits their transparency. Stated differently, tracing the exact computational pathway a network follows to reach a specific prediction is highly challenging, which has led to DL frameworks being widely characterized as "black boxes" [49–51]. To mitigate this limitation, the subfield of Explainable Artificial Intelligence (XAI) has emerged to improve the interpretability of deep neural networks. Ref. [50] classified XAI methodologies into two primary categories: intrinsic and post hoc. The authors explained that intrinsic frameworks are inherently transparent, making their internal decision-making mechanics straightforward to comprehend. Conversely, post hoc methodologies operate by illuminating the specific contextual elements of the input features that drove a classification decision within an already trained model. These tools typically leverage visual analytics, local explanations, and simplification techniques, among other functionalities.

Ref. [52] argued that integrating XAI within clinical environments can assist medical experts by delivering diagnoses that are transparent, comprehensible, and explainable. Successful clinical translation, however, requires that ML models possess the capacity to explicitly justify their predictive logic. The authors further emphasized that the proper execution of XAI strategies is vital for increasing clinical trust and accelerating the adoption of these tools in medicine. By designing three distinct explainability strategies to support human decision-making in clinical settings, they concluded that refined XAI models could become foundational instruments in the physician's diagnostic workflow.

This perspective is reinforced by [49], who argued that the black-box nature of most DL predictions leaves them devoid of interpretability regarding how an outcome was generated. Consequently, healthcare professionals find it difficult to understand or fully rely on the outputs produced by classical ML models, which restricts their deployment within public health ecosystems. Building clinical trust requires incorporating systematic explainability and interpretation layers directly onto model outputs. Implementing these safeguards could position DL frameworks as a compelling and reliable choice for optimizing public health and patient outcomes.

In this investigation, we utilize Local Interpretable Model-agnostic Explanations (LIME) to provide clear interpretability and deliver crucial, actionable insights for clinical decision-making [51]. This XAI methodology deduces explanations by evaluating the behavioral patterns of the complex classifier within the immediate mathematical proximity of the specific case being evaluated, relying on simpler local surrogate models like logistic regression or decision trees [52]. Ref. [49] noted that XAI outputs can be viewed as feature-importance indicators that describe the internal mechanics of a highly complex predictive network.

Accordingly, LIME simplifies the user's capacity to understand the reasoning behind an automated decision by highlighting the most significant and contextually relevant features driving a given prediction. The algorithm monitors changes in model predictions by perturbing the input data. Crucially, rather than trying to explain the behavior of the entire model across the global dataset, LIME generates localized explanations for individual, isolated cases using unseen testing partitions. It leverages individual feature prediction probabilities to explain the rationale behind a specific output [49]. Visualizing the explanations produced by these XAI tools makes it easier to generalize and communicate findings within clinical environments, thereby increasing the reliability of DL tools and accelerating their acceptance in medicine.

The current study evaluates the deployment of SEDL architectures to ascertain CoDs from unstructured VA text narratives. At present, there is a paucity of research exploring the intersection of XAI and SEDL configurations, particularly within the verbal autopsy domain. While one notable investigation [42] utilized XAI methods, their framework did not incorporate stacked ensemble designs. The contribution of this paper is to advance current healthcare-focused XAI research by developing highly transparent and interpretable DL pipelines. These models are designed to be economical, efficient, accurate, consistent, and error-free, ultimately improving the operational quality of VA reporting.

Here, we apply LSTM, CNN, and SEDL networks to achieve easily interpretable CoD classifications from VA records, clarifying how each model translates text inputs into predicted outcomes. These explainable DL systems can be seamlessly embedded into mortality surveillance workflows to augment manual human annotation and interpretation, offering substantial utility to clinicians and public health researchers.

Materials and Methods

This section details the methodologies and analytical frameworks applied throughout this investigation.

Study design

This investigation utilizes a retrospective cross-sectional design driven by secondary data analysis. Data standardization and normalization procedures were executed within a Python Spyder environment. The resulting data frames—comprising cleaned VA narratives, model performance indicators, and classified VA categories—were systematically structured and hosted within a PostgreSQL version 4.2 Relational Database Management System.

Study population

The empirical data for this study were obtained from the Agincourt Health and Demographic Surveillance System (HDSS). Established in 1992, the Agincourt HDSS is situated in a rural region of northeastern South Africa, specifically within the Bushbuckridge sub-district of the Ehlanzeni District in Mpumalanga Province. The geographical footprint of the Agincourt study area spans roughly 420 km². Based on the 2019 Agincourt census profile, the baseline population stood at 116,247 residents distributed across 28 villages and 22,716 households. This cohort included 55,961 males, 60,280 females, 11,724 children under the age of five, and 928 school-age youth between 5 and 19 years old [53].

Data source and description

This study leveraged the longitudinal Agincourt HDSS registry encompassing records from 1993 to 2015. Our DL models ingested textual VA narratives to classify CoDs across twelve distinct pathological categories established in historical literature [3, 9, 54, 55]. The source matrix contains 287 feature columns and 16,338 individual patient records. In our modeling framework, all documented symptoms along with the raw VA narrative texts served as the predictor matrix X, while the validated physician diagnoses—coded according to the International Classification of Diseases 10th Revision (ICD-10)—constituted the target vector Y across all 16,338 observations.

Table 1 outlines the specific tracking labels assigned to each of the twelve disease groups evaluated in this study.

Table 1. Labelling of the twelve disease classes.

Disease Group	Sample Count	Class Identifier
HIV/TB-Related Conditions	3388	0

Other Infectious Diseases	964	1
Metabolic Disorders	242	2
Cardiovascular Diseases	140	3
Indeterminate Causes	1468	4
Maternal and Neonatal Conditions	121	5
Abdominal Diseases	117	6
Neoplastic Disorders	93	7
Neurological Disorders	57	9
Respiratory Diseases	46	10
Other Non-Communicable Diseases (NCDs)	21	11

Legend: HIV—human immunodeficiency virus, TB—tuberculosis, NCD—non-communicable diseases.

Preprocessing

The preprocessing pipeline targeted incomplete, noisy, and discordant data points using systematic data cleaning, munging, transformation, and dimensionality reduction techniques [56]. Text cleaning operations standardized the narrative inputs and removed extraneous characters that offered no semantic value to the models. This was achieved by converting all text strings to lowercase and stripping out punctuation marks, whitespaces, digits, and special symbols.

Spelling anomalies were corrected using the TextBlob library in Python, and tokenization was applied to split the raw text blocks into discrete word tokens. Irrelevant lexical elements were filtered out using the NLTK library's standard English stopwords list. Lemmatization was then carried out to reduce word variations to their base dictionary form, or lemma. Missing or null values were removed entirely from the dataset to eliminate downstream modeling bias. To ensure uniform feature scaling, the Python StandardScaler was applied to normalize the magnitudes of our predictors X and target outputs Y within an equivalent range. Finally, to counteract severe class imbalances identified during exploratory data analysis, the Synthetic Minority Oversampling Technique (SMOTE) was applied exclusively to the training dataset.

Word embedding and representation

This study opted for Global Vectors for Word Representation (GloVe) to manage distributed word embeddings, with the objective of mitigating the architectural challenges of high dimensionality by mapping text features into a low-dimensional vector space. Distributed word vectors serve as the foundational processing tier of a DL network, operating on the principle that words with close semantic meanings share comparable contextual distributions.

The GloVe framework was chosen over traditional Term Frequency (TF) and Term Frequency-Inverse Document Frequency (TF-IDF) models because of its robust capacity to capture latent semantic relationships. In contrast, standard TF frameworks are frequently dominated by high-frequency words, and TF-IDF models often fail to detect deeper contextual similarities between different terms. The mathematical properties of the GloVe algorithm are comprehensively detailed in [57, 58].

Text classification with deep learning architectures

Text classification tasks in this investigation were carried out using Long Short-Term Memory (LSTM), CNN, and SEDL architectures. Unlike standard distributed word vectors, these architectures are structurally equipped to handle extensive vocabularies containing out-of-vocabulary or unknown terms (words that appear intermittently within testing sets but are absent from the training corpus). Within existing literature, these configurations are recognized as character embedding models because they interpret each word as a sequential arrangement of its constituent characters [59].

Long short-term memory

The LSTM network represents a specialized variant of the recurrent neural network (RNN) that integrates feedback loops between its processing layers, providing it with dynamic temporal characteristics [60]. Standard RNN structures are designed to process sequential information by executing uniform operations across each successive token in a sequence. These models ingest current input strings while simultaneously referencing hidden states derived from prior computational steps [61], utilizing internal memory components to process sequential tasks. Generally, RNN configurations are built from three primary structural tiers: an input layer, hidden processing layers, and an output layer [60].

However, basic RNN architectures frequently suffer from the vanishing gradient problem when attempting to model long-range dependencies. This occurs when the gradient values used to update the network weights decrease exponentially toward zero during backpropagation, undermining overall accuracy and making hyperparameter tuning difficult. The structural properties of basic RNNs are described in [59].

LSTM networks overcome the vanishing gradient limitation through a gated architecture comprising three functional elements: a forget gate, an input gate, and an output gate. This design makes the LSTM exceptionally efficient relative to standard RNNs, as it can selectively purge extraneous data from its memory cell, write new target features to memory, and isolate the immediate contextual attributes most relevant to the task [61]. The inclusion of the forget gate differentiates the LSTM from basic RNN frameworks, enabling stable backpropagation across extended sequences. The network determines its current hidden state by calculating and balancing the interactions across these three internal gates [59]. Figure 1 illustrates the structural comparison between the LSTM deep learning framework and a conventional RNN.

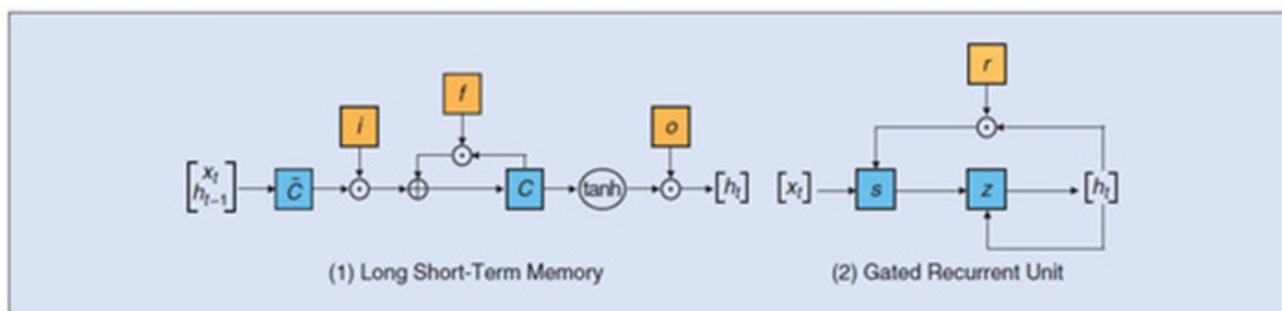


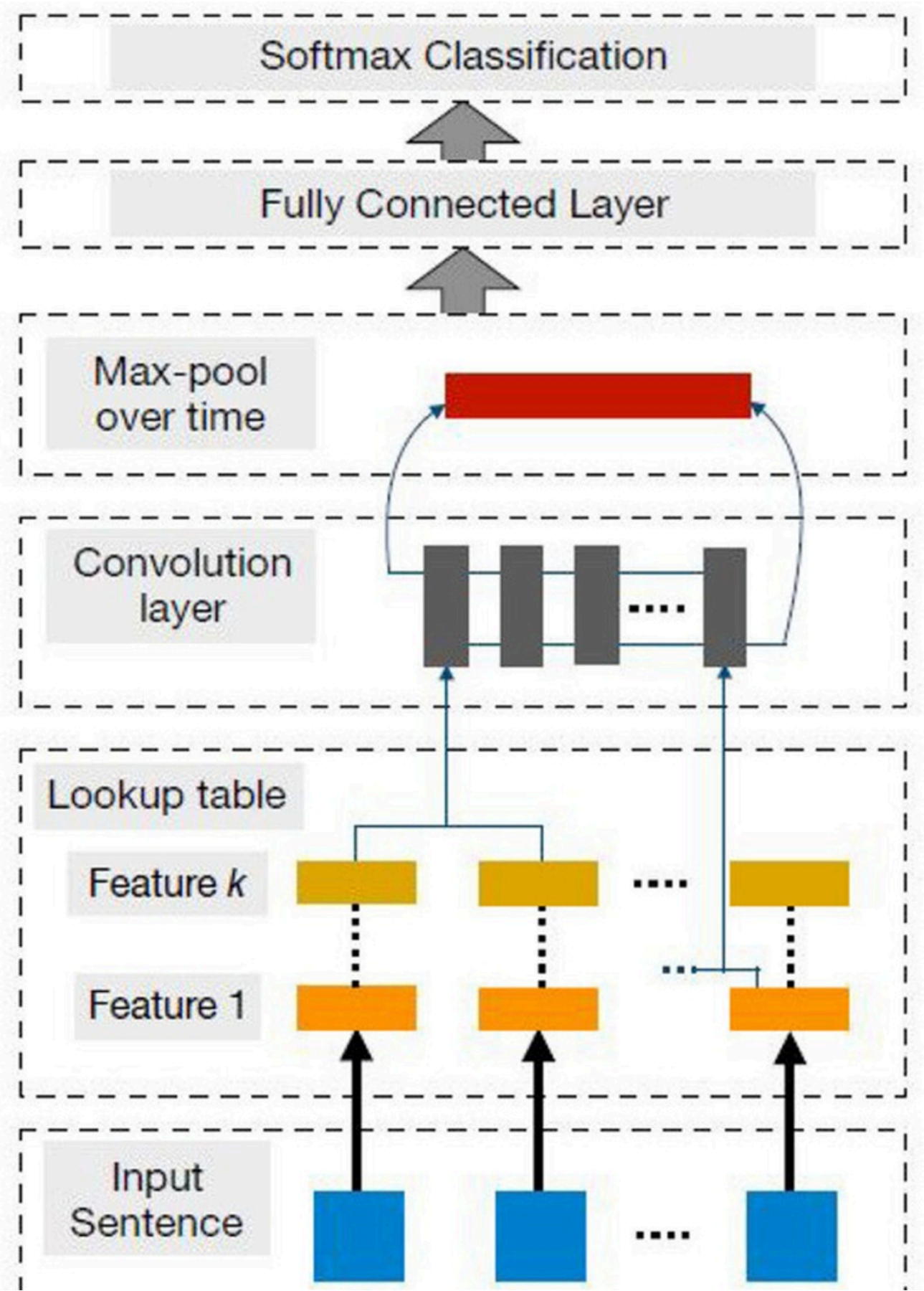
Figure 1. LSTM deep learning framework. Source: [59].

Convolutional neural network

The Convolutional Neural Network (CNN) was originally adapted as a feature extraction mechanism intended to capture high-level features from n-grams, meaning that these extracted attributes are highly applicable across diverse NLP tasks [59, 62, 63]. The CNN operates under a strict weight constraint, which dictates that an individual convolutional layer must apply identical weights across the inputs for all processing nodes residing within that specific layer. Consequently, this architectural constraint makes a CNN significantly more efficient to train than a standard fully connected neural network [61].

The primary layers of a CNN comprise an input layer, convolutional layers, pooling layers, fully connected layers, and a final softmax function to execute the classification task. The underlying mechanics of the CNN framework are described in

[59]. **Figure 2** illustrates the pictorial representation of the CNN architecture.



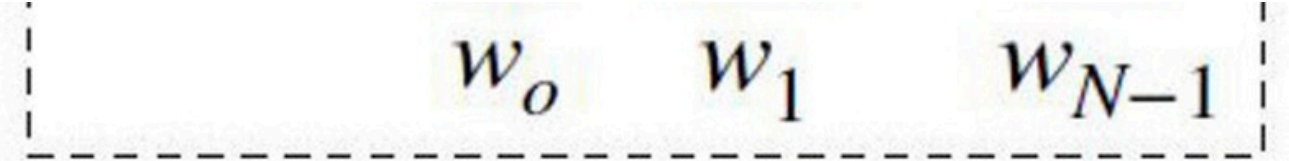


Figure 3 presents a diagrammatic layout of the ensemble DL techniques utilizing a stacking approach. Our implementation incorporates multiple individual base learners (B_1 – B_n) that process data through several successive iterations ($Level_1$ – $Level_n$) and subsequently channel their intermediate outputs into a meta-learner (M_{output}) to produce the final model prediction. This investigation leveraged the structural stacking principles formulated in the work of [64].

The underlying dataset was initially partitioned into n equal components, where each n^{th} -fold cross-validation segment was isolated for testing while the remaining sections were utilized to train the models. We systematically generated pairs of train–test sub-datasets to produce diverse intermediate model predictions, which were then supplied as structural inputs into our meta-model for the definitive classification. Accordingly, this multi-tiered approach boosts overall model performance while simultaneously mitigating the potential for systemic bias within the final outcomes [47].

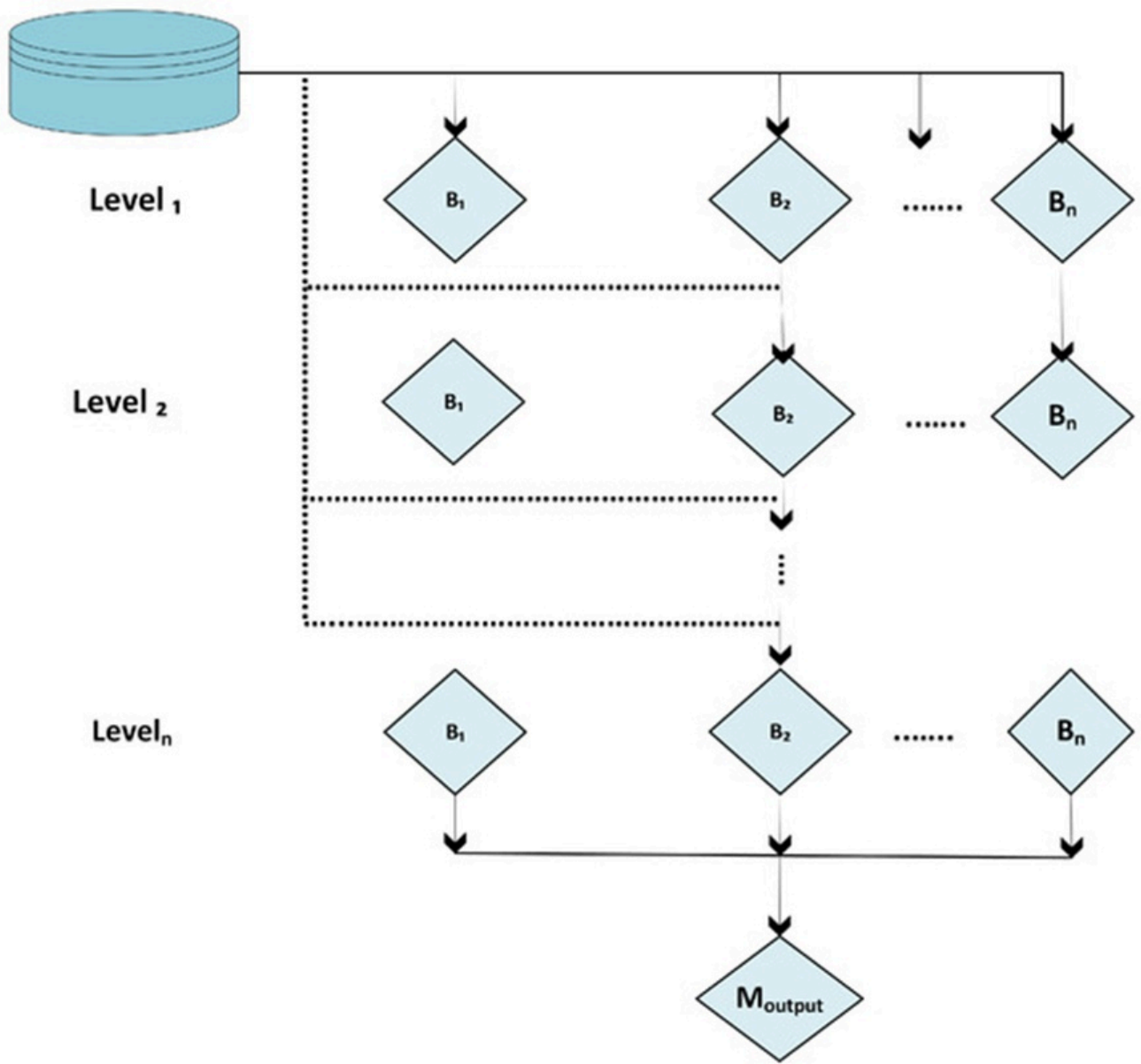


Figure 3. Stacking of our models. Source: [47].

Explainable AI using LIME

Figure 4 provides a visual representation of the general schematic pipeline detailing the integration of XAI via post hoc explanations using a simplification method, namely LIME. Initially, the unstructured VA narratives are supplied as input vectors into the black-box SEDL models. Following this classification phase, the model output is interpreted using the LIME component to supply both the explicit CoD diagnosis and its corresponding explanation to medical experts. In this work, we implemented the mathematical XAI framework (LIME) as reported in [50, 52].

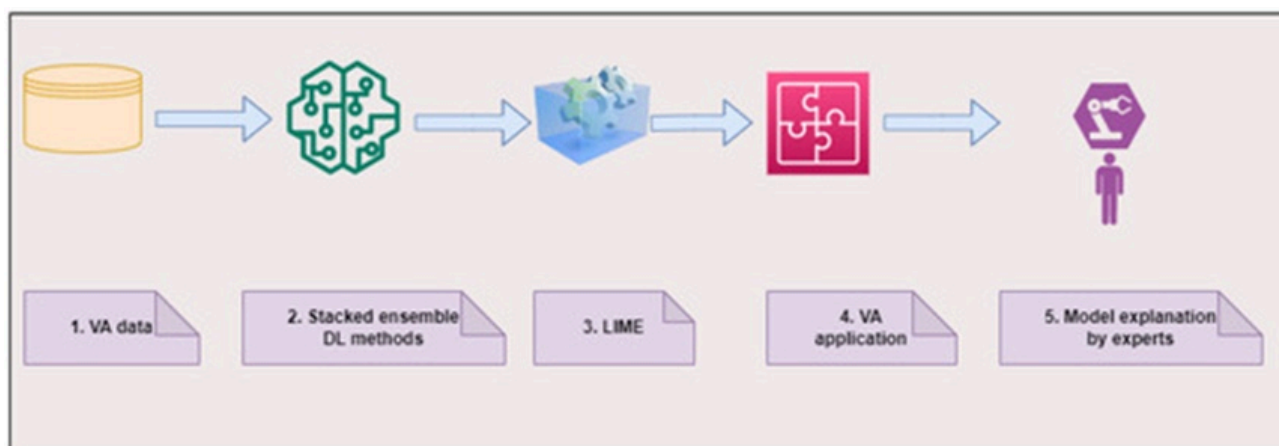


Figure 4. General pipeline of XAI implementation using VA data.

Experiments: Deep Learning Models

This section details the distributed word vector configurations, the specific DL architectures, and the Python software libraries deployed throughout this study. Global Vectors (GloVe) were utilized to generate vectorized word representations via the Keras Python library, which we subsequently re-trained using our localized study corpus. We constructed a 128-dimensional vector space for each word token by utilizing a window size of 5. The input VA narratives were uniformly padded to ensure a consistent sequence length of 250 words across all text records. During empirical testing, it was observed that scaling the vector dimensions to larger values yielded no statistically significant impact on performance.

The implemented CNN architecture comprised four distinct 1D convolutional layers with channel depths of 128, 256, 256, and 512, which were interleaved with corresponding max-pooling and dropout layers. The kernel size for both the convolutional and max-pooling operations was set uniformly at 3. A flattening layer was integrated to compress all feature maps generated by the convolutional steps into a single, continuous feature vector, which was then linked to the fully connected layer. This fully connected layer was configured with 128 hidden units and also featured a dropout layer.

All internal network layers utilized the rectified linear unit (ReLU) activation function, save for the final classification layer, which utilized a softmax function. For model optimization, cross-entropy loss was selected as the objective metric, and Adam was utilized as the primary optimization algorithm. The network was configured with an early stopping routine to halt training if no further performance gains were detected.

The RNN pipeline, implemented via the LSTM variation, incorporated two Gated Recurrent layers containing 256 and 512 nodes, respectively. The ultimate fully connected layer contained 12 distinct nodes at its termination and was paired with a dropout layer configured at a rate of 0.002. The final classification layer employed a softmax activation function, and model optimization was conducted by measuring cross-entropy loss and leveraging Adam as the optimization algorithm.

This framework was likewise paired with an early stopping function to terminate training if the model failed to achieve further effective training steps.

We systematically tuned both batch size and the number of epochs as components of batch normalization, aiming to standardize the distribution across each input dimension of our processed dataset [65]. Furthermore, we adjusted the optimization algorithms used to train the network, evaluating each under its default parameter baselines. Additionally, we optimized the learning rate—which dictates the step size for updating network weights at the conclusion of each batch—and the momentum parameter, which determines the degree to which previous weight updates influence the current weight modification.

Optimization routines were also applied to select the network weight initialization strategy by benchmarking all available computational techniques. Moreover, we fine-tuned the activation functions that govern the non-linear transformations of individual neurons and dictate their firing thresholds. The dropout rate, combined with weight constraints, was optimized by adjusting the dropout percentage for regularization to limit overfitting and enhance the model's capacity to generalize to unseen datasets. Finally, we adjusted the total number of neurons within the hidden layers, recognizing that larger network Topologies demand extended training times and that batch size and epoch limits should ideally be optimized in tandem with hidden unit counts.

Our standard individual DL architectures failed to converge, as they were unable to achieve our target optimal performance benchmark of 80% established in our introductory research reported in [21]. Consequently, we transitioned our experimentation toward ensemble DL strategies, specifically deploying SEDL architectures. A sequential model from the Keras ML library was constructed, consisting of one global average pooling layer followed by four dense layers configured with 1024, 512, 128, and 64 units, respectively.

Additionally, to handle architectural complexity and prevent overfitting, we integrated dropout layers and implemented structural regularization using the Lasso (L1) technique. We selected Adam as our optimization algorithm, designated a patience parameter of 3, and configured the architecture to run for up to 500 epochs with a batch size of 128. This framework served as the foundation for our base learners, which comprised an ensemble of 100 member models.

DL Model Evaluation

In this study, we utilized optimization and performance learning curves to assess the efficacy of our DL models. Specifically, we used training and validation datasets to visualize cross-entropy loss and model accuracy across multiple successive iterations. The optimization learning curves plot both training loss and validation loss over elapsed training epochs.

The training loss curve illustrates the degree to which a model fits the underlying training observations, whereas the validation loss curve reveals how effectively the model adapts to entirely new, unseen data points. Parallel to this, performance learning curves—which leverage accuracy as the primary evaluation metric—were implemented. Here, training accuracy indicates the model's fit on the training data, while validation accuracy tracks the model's generalized capacity to classify new, independent records. The operational principles of these learning curves are described in detail within the studies of [61, 66, 67].

We also report on the total accuracy achieved by our architectures. Accuracy measures the proportion of correctly predicted results across all evaluated categories relative to the total number of samples [18, 20]. True Positives (TPs) and True Negatives (TNs) represent the count of classification outcomes where our predictive system correctly identifies positive and negative instances, respectively.

In our experimental framework, a TP denotes a predicted positive VA narrative that maps to a specific disease category among the twelve available classes and is clinically confirmed as positive. Conversely, a TN denotes a predicted negative

VA narrative for a specific disease class that is confirmed as negative. Alternatively, False Positives (FPs) and False Negatives (FNs) represent the counts of classification outcomes where our models erroneously predict the positive or negative classes.

In this context, an FP indicates a VA narrative that the model flags as positive for a certain disease category when it is actually negative, whereas an FN represents a narrative that the model flags as negative for a disease category when it is actually positive. The mathematical formulation for calculating accuracy is expressed as follows.

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (1)$$

Results and Discussion

In this section, we analyze the outcomes delivered by our DL architectures across the classification of VA reports into twelve distinct medical categories.

Our initial data cleaning involved removing 2247 VA reports containing null records, leaving 14,091 text cases to serve as input for the DL frameworks. Additionally, 5170 missing target outcome values were excluded from the analysis. Within the dataset, the maximum narrative length spanned 715 words (comprising 3949 characters), whereas the shortest text consisted of seven words (comprising 55 characters). Across the entire corpus, the narratives exhibited an average sentence length of 99.76 words.

To address class imbalances, SMOTE was utilized to generate twelve uniform disease classes, each containing 3388 samples for the training partition. This procedure was necessary because prominent classes originally held a significantly larger volume of data points relative to underrepresented minority classes. Systems that overlook proper data balancing protocols frequently exhibit architectural bias, yielding substantial misclassification errors and distorted analytical insights.

Every individual VA narrative was treated as a discrete, short text document composed of a sequence of terms. These word sequences, processed as unigrams ($n = 1$) and bigrams ($n = 2$), served as the input features for our DL structures to forecast the most probable CoD among the twelve target disease categories. Our objective focused on projecting these target feature values onto the unseen test partition. To facilitate mathematical processing within our predictive models, these text sequences were converted into dense, numeric vector spaces via Global Vectors (GloVe) word embeddings.

Model performance

Table 2 outlines the performance evaluation metrics derived for the three DL systems (LSTM, CNN, and SEDL) using optimization routines and machine learning performance curves. The SEDL framework demonstrated superior performance over both the LSTM and CNN variants in classifying diseases. Based on the ML curves, SEDL reached a training accuracy of 82.1%, whereas the LSTM and CNN achieved 76.11% and 76.35%, respectively. Regarding validation accuracy, SEDL achieved 82%, outperforming the LSTM at 67.05% and the CNN at 66.16%. On the evaluation test set, the observed accuracies were 82% for SEDL, 67% for LSTM, and 66.2% for the CNN. Test loss metrics reached 1.15% for SEDL, 11.95% for LSTM, and 12.64% for the CNN. Correspondingly, SEDL, LSTM, and the CNN demonstrated overall test accuracy rates of approximately 82%, 67%, and 66.2%, indicating that the correct target class was assigned to these exact percentages of test samples by each respective model.

Table 2. Performance evaluation of models.

Model	Test Loss (%)	Test Accuracy (%)	Validation Accuracy (%)	Training Accuracy (%)
LSTM	11.95	67	67.05	76.11
CNN	12.64	66.2	66.16	76.35
SEDL	1.15	82	82	82.1

The generalization capability of the DL frameworks on novel, unseen data is illustrated through the learning curves in **Figures 5 and 6**. The trajectories demonstrate that the architectures successfully generalize, enabling them to construct reliable predictions on new instances by capturing the core statistical properties of the training data. To safeguard generalization and mitigate the risks of overfitting or underfitting, architectural complexity was regulated by incorporating a dropout rate of 20%. Furthermore, regularizing mechanisms including early stopping (configured to monitor validation loss), a patience threshold of 3 epochs, and a minimum delta variation of 0.001% were integrated to halt training when performance improvements stagnated. The minimal discrepancy observed between the validation and test accuracies reflects a highly robust model fit.

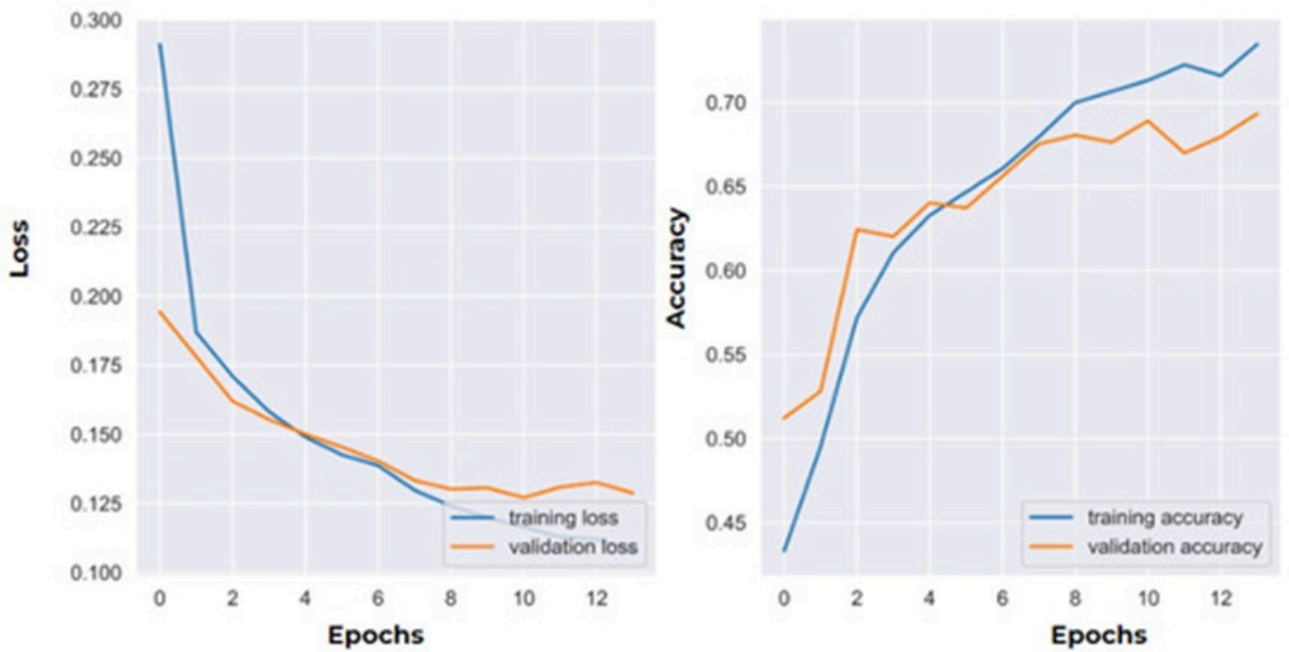


Figure 5. Learning curve depicting performance evaluation of CNN.

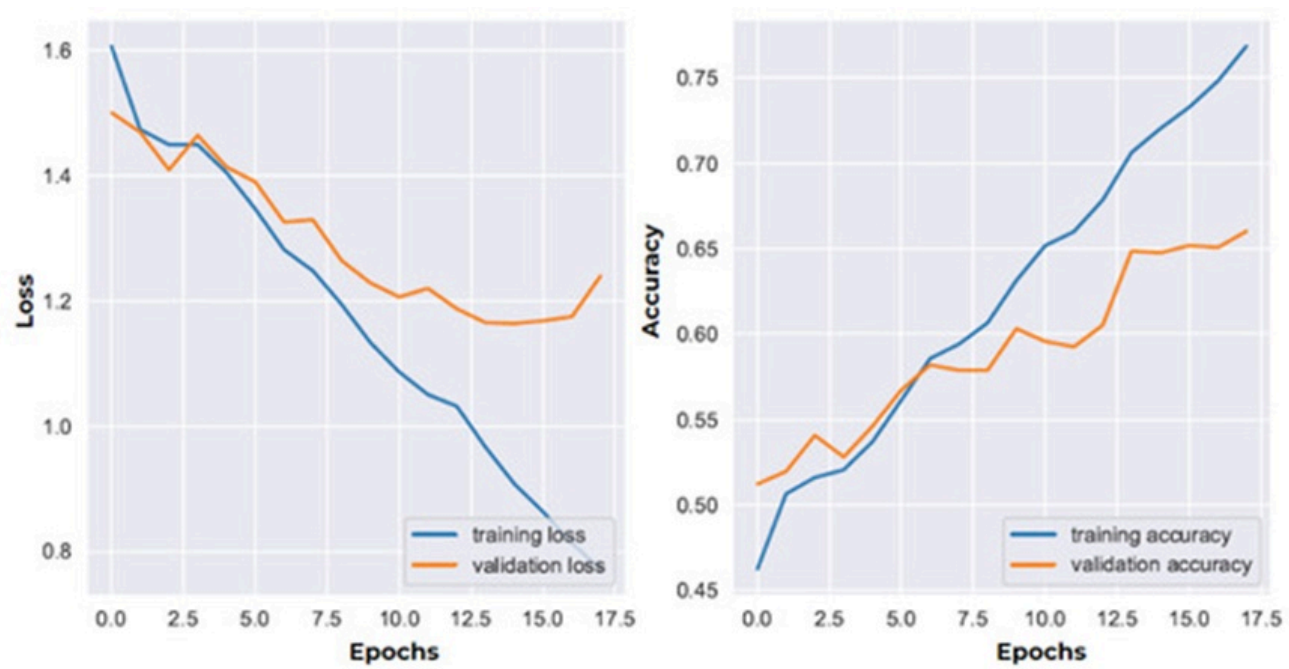


Figure 6. Learning curve depicting performance evaluation of LSTM.

The training and testing trajectory behaviors of the stacked ensemble network are illustrated in **Figures 7 and 8**. The narrow margin between the training and testing performance curves indicates that the ensemble architectures successfully avoided both underfitting and overfitting states.

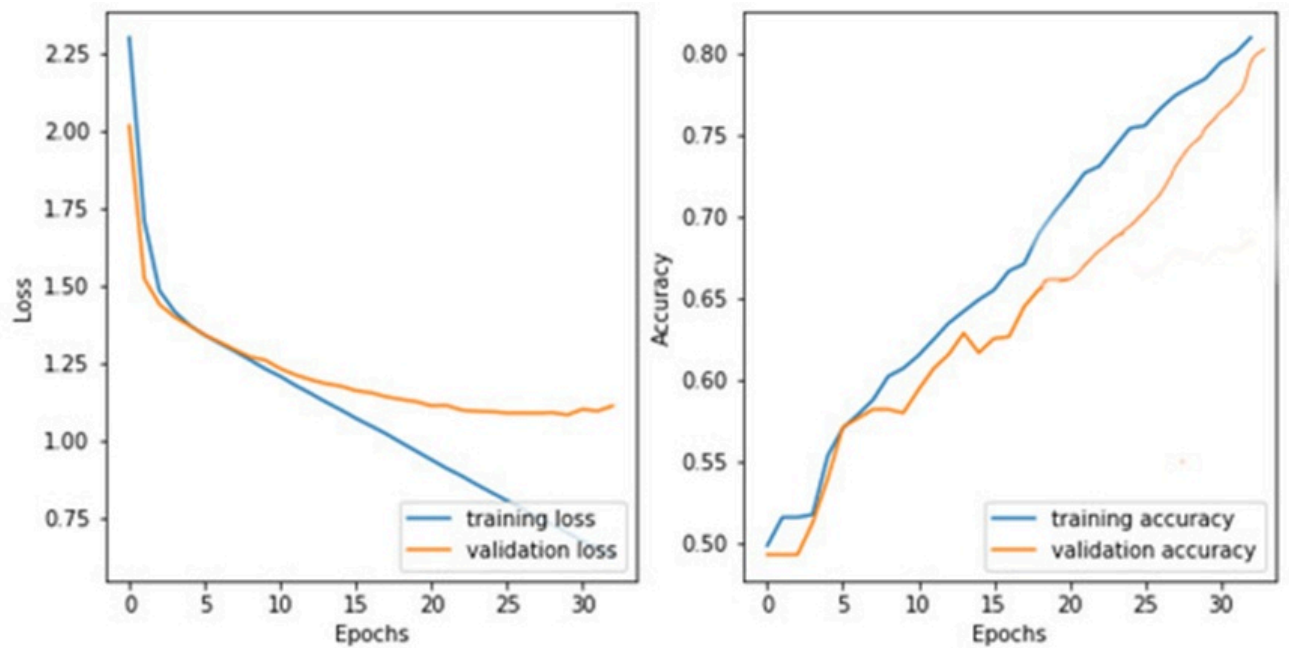


Figure 7. Learning curve depicting SEDL performance evaluation during model training.

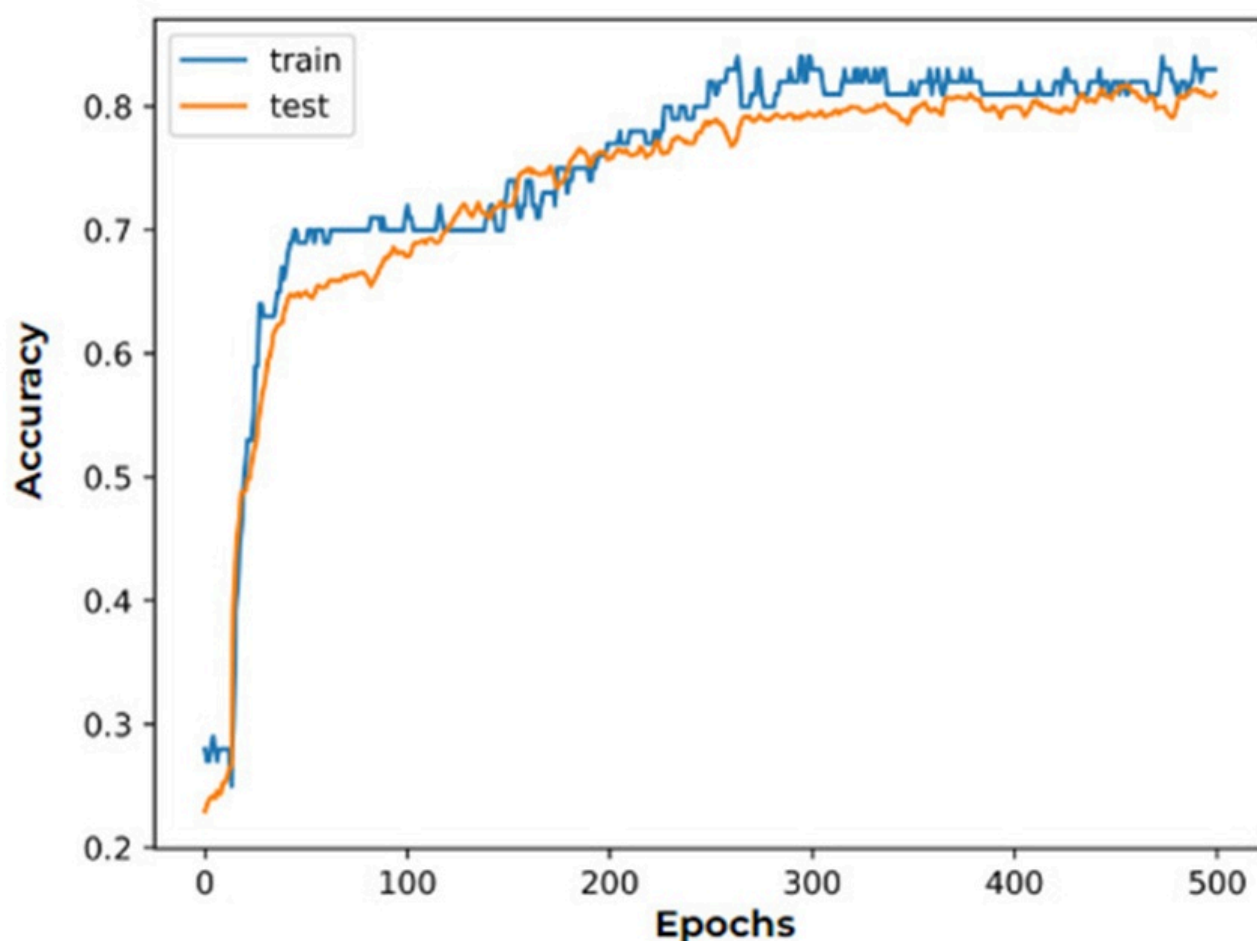


Figure 8. Learning curve depicting SEDL performance evaluation.

Model explainability

This subsection details the interpretability analysis conducted on our SEDL structures using the LIME framework. While our models produced varying CoD predictions and explanations across cases, deeper diagnostic clarity was established by leveraging LIME to improve transparency.

Figures 9 and 10 illustrate text narratives involving metabolic, maternal, and neonatal diagnoses that the SEDL framework classified correctly. In an instance where a patient passed away from metabolic diabetes, the narrative highlighted specific symptoms like "high sugar" and "sugar diabetic"—terms that subsequently registered high probability weights within our XAI LIME model. Another successful classification is presented in **Figure 9**. Notably, LIME directed the prediction framework based on the clinical narrative surrounding the patient's death by prioritizing key terms such as "baby", "incubator", "mother", "delivery", and "birth", all of which carried high feature probabilities. It is apparent that in multiple cases, the models deduced the appropriate CoD by establishing patterns among reported symptoms, clinical treatments, and diagnostic laboratory evaluations. These comprehensive descriptions consistently corresponded with higher predictive accuracy across our models.

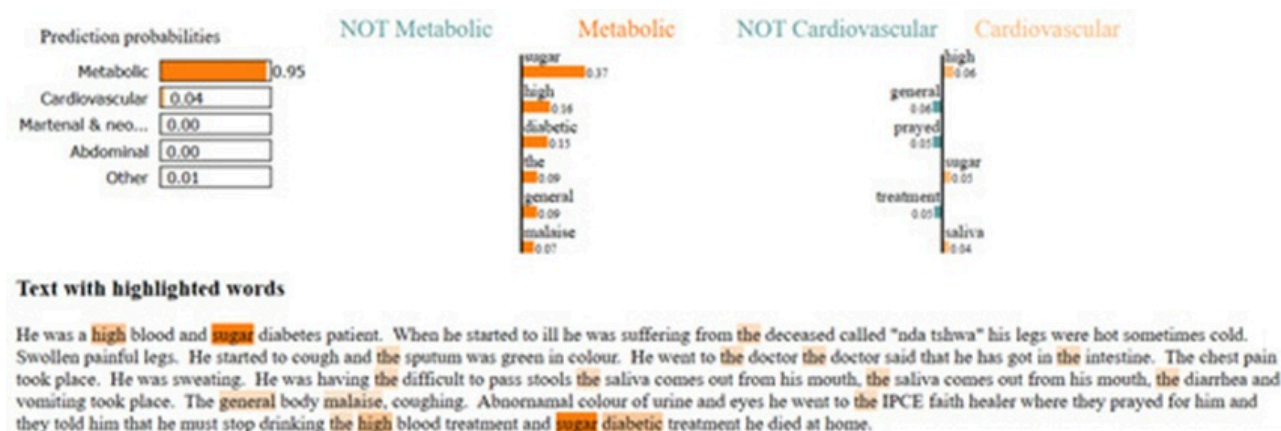


Figure 9. Example of LIME correct prediction.

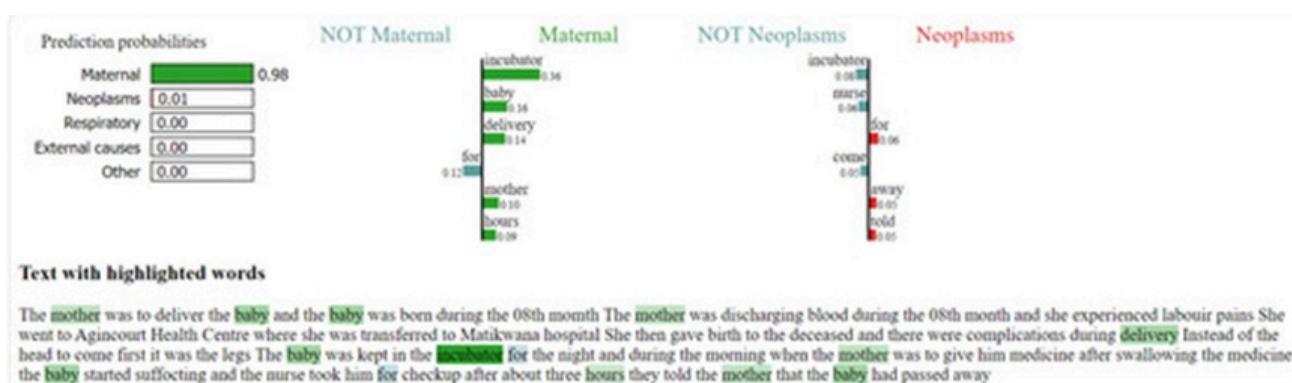


Figure 10. Example of LIME correct prediction.

Conversely, **Figures 11 and 12** display instances of incorrect classifications, where the architectures erroneously predicted HIV and TB as the primary CoDs, whereas the established gold standard diagnoses were non-communicable diseases and respiratory conditions, respectively. Diagnostic errors often occurred due to limitations in NLP processing proficiency, specifically when narratives contained extraneous historical background details that had no functional relevance to the immediate CoD. Consequently, accurate model predictions depended on prioritizing the direct clinical events leading to death rather than distal historical patient records, an issue clearly visible in the text segment of **Figure 12**.

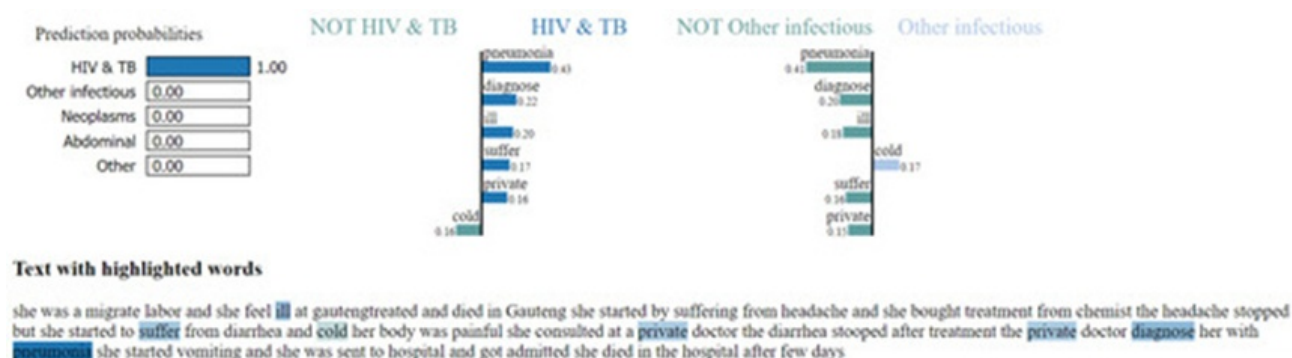


Figure 11. Example of LIME incorrect prediction.



Figure 12. Example of LIME incorrect prediction.

We executed an experimental trial designed to gather mean opinion scores (MOSs) from medical practitioners to evaluate the clinical validity of the LIME explanations and model classifications. This process aimed to verify whether the rationales generated by LIME aligned with clinical reasoning. Ten physicians reviewed an identical set of 20 randomly sampled LIME predictions. The qualitative assessment categories and their corresponding numerical weights were defined as follows:

1. Bad.
2. Poor.
3. Fair.
4. Good.
5. Excellent.

The final computed MOS represents the collective perceived quality of the LIME explanations. The evaluation yielded a high MOS of 4.21, demonstrating that the interpretive justifications produced by LIME were clinically logical and coherent.

Determining the cause of death (CoD) from verbal autopsy (VA) records using text narratives alone is traditionally a manual, labor-intensive, error-prone, and expensive process. Specifically, the prolonged turnaround time required for medical professionals to review these written accounts and establish an accurate diagnosis remains a persistent bottleneck. Despite ongoing initiatives to refine VA systems, the data extraction phase is constrained by several limitations that delay CoD determination. This lag directly hinders timely reporting, which is essential for updating civil registration infrastructures and reinforcing public health priorities.

In this research, we deployed innovative deep learning (DL) architectures exclusively on VA narratives to minimize the time physicians spend reviewing records before rendering a diagnosis. While existing literature primarily relies on the structured, closed-ended responses within VA questionnaires for feature extraction, our work demonstrates that the unstructured narratives contain rich, high-value clinical insights [3]. In contrast to standard methodologies that rely on closed-ended question fields to develop predictive models, our findings show that utilizing text narratives alone generates comparable classification outcomes that align closely with physician-certified CoDs. Furthermore, to maximize the utility of VA narratives, it is essential to adopt modern technical approaches, such as using digital tablets to record interviews with the families of the deceased and utilizing automated transcription systems. Transitioning away from condensed interview summaries toward complete verbatim transcriptions could yield substantially more reliable data, driving error rates below 6% [42].

Here, LSTM networks and CNNs were applied to classify CoD solely from verbal accounts. In line with previous research, we observed that DL techniques do not invariably deliver optimal performance out of the box, given their heavy reliance

on massive datasets and extensive computational infrastructure. Consequently, this study encountered obstacles regarding constrained training sample sizes alongside a pronounced occurrence of out-of-vocabulary terms and high lexical variability (words carrying broad, context-dependent meanings within the medical domain) [42]. Although both basic DL models underperformed relative to expectations, the LSTM architecture achieved superior metrics compared to the CNN.

Nevertheless, the lower performance of the LSTM suggests it was less effective at handling the sequential nature of these specific narratives and struggled to fully overcome the vanishing gradient problem. This implies that during gradient-based model training, the gradient values diminished significantly, which kept the network weights static or unadjusted and prevented effective learning. This specific limitation likely impaired the utility of the LSTM's forget gate mechanism (the component responsible for discarding historical computations from memory, conserving only high-utility data, and focusing exclusively on recent, contextually relevant memory inputs), which explains the observed performance limits.

Similarly, the CNN delivered only moderate performance. The training behavior indicates that the model did not effectively enforce weight sharing across all nodes within each respective convolutional layer. Additionally, the configuration lacked an optimal number of hyperparameters at each layer to ensure a balanced, efficient learning process across the available data points. Furthermore, this architecture failed to fully capitalize on its capacity to feed integrated outputs from earlier layers into subsequent processing stages, an ability that is typically crucial for uncovering highly abstract and intricate feature relationships.

As highlighted by [31], DL methodologies require extensive training datasets to reach peak classification accuracy; when data points are sparse, these models struggle to properly evaluate feature weights and isolate the highly discriminative patterns required for reliable prediction. Conversely, our foundational research documented in [21] demonstrated that conventional machine learning approaches outperform basic DL configurations when analyzing VAs. To move past these basic DL limitations, we developed and investigated SEDL frameworks.

Our SEDL architectures delivered optimized outcomes compared to standard DL frameworks, successfully exceeding the 80% performance benchmark established in our previous study [21]. This improvement occurred because the stacked ensemble configuration leveraged iterative model training to combine the unique strengths of each individual base classifier. These empirical results indicate that SEDL configurations can be deployed effectively for CoD determination using text narratives alone.

Because DL models inherently function as uninterpretable "black boxes," establishing them as viable alternatives to the physician-certified gold standard remains challenging. The public health sector routinely exhibits skepticism toward automated systems for disease diagnosis and prognosis, given that this high-stakes domain demands highly accurate models that offer transparent interpretability. However, the outcomes of this study position our SEDL framework as a compelling alternative to manual physician reviews, matching observations reported by [42]. Automated methodologies naturally reduce human errors and subjectivity, thereby promoting clinical standardization. Additionally, machine learning architectures scale efficiently to process and analyze massive datasets that would overwhelm human resources. Automated frameworks are also highly time-saving (minimizing narrative evaluation periods), cost-effective (requiring fewer resources than manual reviews), immune to human fatigue, and consistent in CoD assignment (as iterative model training continuously refines diagnostic and prognostic precision).

Gathering accurate information from the relatives of a deceased individual remains a significant barrier within the VA framework due to the subjective nature of recall bias. The findings of this research offer an opportunity to increase systemic efficiency by automating text analysis, reducing the manual labor and time required for human review, and ultimately enhancing VA reporting and clinical decision-making.

It is important to emphasize that clinical datasets are not always available in massive volumes; when sample sizes are small or constrained, conventional ML strategies remain preferable. A comparative review between the results achieved here and our prior study using conventional ML techniques [21] confirms that traditional algorithms deliver superior classification performance when applied to small sets of VA narratives. Nevertheless, the core novelty of this research lies in implementing SEDL systems that are made readily interpretable through the integration of LIME. Moreover, the validation metrics derived from physician reviews present this framework as a highly capable, accurate, and economical alternative, fostering greater institutional trust in automated public health applications.

This study was constrained by the limited volume of training samples available, which are critical for maximizing the accuracy of DL architectures. Furthermore, the high degree of clinical lexical variability restricted overall model performance. The analysis was also limited by the duration of the recall window, which may have introduced recollection bias into the collected VA records. Additionally, variations in training levels or interviewing approaches among data collectors could have introduced systematic bias during data collection. Finally, the VA dataset lacked native linguistic terms, which may restrict the broader generalizability of the findings and limit further optimization of the predictive model.

Conclusion

This investigation demonstrates that our SEDL architecture can be successfully integrated into mortality cause determination processes alongside human coding and review workflows. Consequently, these insights offer a baseline for establishing greater trust in algorithmic diagnosis and prognosis within public health domains. Adopting this framework can ultimately decrease the turnaround time for CoD classification, lower operational costs, reinforce civil registration frameworks, support health resource allocation, and guide policy formulation. Future research initiatives will focus on utilizing record linkage methodologies to track under-reported morbidity trends by systematically matching HDSS population records with NHLS public health data registries.

Abbreviations

The following abbreviations are used in this manuscript:

AI	Artificial Intelligence
CCVA	Computer-Coded Verbal Autopsy
CoD	Cause of Death
CNN	Convolutional Neural Network
DL	Deep Learning
XAI	Explainable Artificial Intelligence
LIME	Local Interpretable Model-agnostic Explanations
LSTM	Long Short-Term Memory
ML	Machine Learning
NLP	Natural Language Processing
PCVA	Physician-Coded Verbal Autopsy
RNN	Recurrent Neural Network
SEDL	Stacked Ensemble Deep Learning
VA	Verbal Autopsy

Acknowledgements

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

Conflict of interest

None

Financial support

None

Ethics statement

The study was conducted in accordance with the Declaration of Helsinki and approved by the Institutional Review Board (or Ethics Committee) of the University of the Witwatersrand Faculty of Health Sciences, Human Research Ethics Committee (Medical), (protocol code M1911132 and approved 23 September 2020).

Received: 14 Oct 2024

Revised: 31 Oct 2024

Accepted: 08 Dec 2024

Published online: 15 March 2025

Rights and permissions

Open Access The author(s) retain copyright. This article is licensed under the Creative Commons Attribution-NonCommercial-ShareAlike 4.0 International License. It may be shared and adapted for non-commercial purposes with appropriate attribution, an indication of changes, and distribution of adaptations under the same license. Third-party material may be subject to separate terms identified in its credit line. View the license at <https://creativecommons.org/licenses/by-nc-sa/4.0/>.

References

- Nichols EK, Byass P, Chandramohan D, Clark SJ, Flaxman AD, Jakob R, et al. The WHO 2016 verbal autopsy instrument: an international standard suitable for automated analysis by InterVA, InSilicoVA, and Tariff 2.0. *PLoS Med*. 2018;15(1):e1002486. <https://doi.org/10.1371/journal.pmed.1002486>.
- Thomas LM, D'Ambruoso L, Balabanova D. Verbal autopsy in health policy and systems: a literature review. *BMJ Glob Health*. 2018;3(1):e000639. <https://doi.org/10.1136/bmjgh-2017-000639>.

Jebblee S, Gomes M, Jha P, Rudzicz F, Hirst G. Automatically determining cause of death from verbal autopsy narratives. *BMC Med Inform Decis Mak*. 2019;19:127.
<https://doi.org/10.1186/s12911-019-0843-1>.

Soleman N, Chandramohan D, Shibuya K. Verbal autopsy: current practices and challenges. *Bull World Health Organ*. 2006;84(3):239-45.
<https://doi.org/10.2471/blt.05.027698>.

Bailo P, Gibelli F, Ricci G, Sirignano A. Verbal autopsy as a tool for defining causes of death in specific healthcare contexts: study of applicability through a traditional literature review. *Int J Environ Res Public Health*. 2022;19(18):11749.
<https://doi.org/10.3390/ijerph191811749>.

Clark SJ. A guide to comparing the performance of VA algorithms. *arXiv*. 2018. arXiv:1802.07807.

Desai N, Aleksandrowicz L, Miasnikof P, Lu Y, Leitaio J, Byass P, et al. Performance of four computer-coded verbal autopsy methods for cause of death assignment compared with physician coding on 24,000 deaths in low- and middle-income countries. *BMC Med*. 2014;12:20.
<https://doi.org/10.1186/1741-7015-12-20>.

James SL, Flaxman AD, Murray CJ. Performance of the Tariff method: validation of a simple additive algorithm for analysis of verbal autopsies. *Popul Health Metr*. 2011;9:31.
<https://doi.org/10.1186/1478-7954-9-31>.

Byass P, Herbst K, Fottrell E, Ali MM, Odhiambo F, Amek N, et al. Comparing verbal autopsy cause of death findings as determined by physician coding and probabilistic modelling: a public health analysis of 54 000 deaths in Africa and Asia. *J Glob Health*. 2015;5(1):010402.

McCormick TH, Li ZR, Calvert C, Crampin AC, Kahn K, Clark SJ. Probabilistic cause-of-death assignment using verbal autopsies. *J Am Stat Assoc*. 2016;111(515):1036-49.
<https://doi.org/10.1080/01621459.2016.1152191>.

Miasnikof P, Giannakeas V, Gomes M, Aleksandrowicz L, Shestopaloff AY, Alam D, et al. Naive Bayes classifiers for verbal autopsies: comparison to physician-based classification for 21,000 child and adult deaths. *BMC Med*. 2015;13:286.
<https://doi.org/10.1186/s12916-015-0521-2>.

Bouille A, Chandramohan D, Weller P. A case study of using artificial neural networks for classifying cause of death from verbal autopsy. *Int J Epidemiol*. 2001;30(3):515-20.
<https://doi.org/10.1093/ije/30.3.515>.

Flaxman AD, Vahdatpour A, Green S, James SL, Murray CJ. Random forests for verbal autopsy analysis: multisite validation study using clinical diagnostic gold standards. *Popul Health Metr*. 2011;9:29.
<https://doi.org/10.1186/1478-7954-9-29>.

Quigley MA, Chandramohan D, Setel P, Binka F, Rodrigues LC. Validity of data-derived algorithms for ascertaining causes of adult death in two African sites using verbal autopsy. *Trop Med Int Health*. 2000;5(1):33-9.
<https://doi.org/10.1046/j.1365-3156.2000.00502.x>.

Mwanyangala MA, Urassa HM, Rutashobya JC, Mahutanga CC, Lutambi AM, Maliti DV, et al. Verbal autopsy completion rate and factors associated with undetermined cause of death in a rural resource-poor setting of Tanzania. *Popul Health Metr*. 2011;9:41.
<https://doi.org/10.1186/1478-7954-9-41>.

Koopman B, Karimi S, Nguyen A, McGuire R, Muscatello D, Kemp M, et al. Automatic classification of diseases from free-text death certificates for real-time surveillance. *BMC Med Inform Decis Mak*. 2015;15:53.
<https://doi.org/10.1186/s12911-015-0178-2>.

Koopman B, Zuccon G, Nguyen A, Bergheim A, Grayson N. Extracting cancer mortality statistics from death certificates: a hybrid machine learning and rule-based approach for common and rare cancers. *Artif Intell Med*. 2018;89:1-9.
<https://doi.org/10.1016/j.artmed.2018.04.001>.

Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Automatic ICD-10 multi-class classification of cause of death from plaintext autopsy reports through expert-driven feature selection. *PLoS One*. 2017;12(1):e0170242.
<https://doi.org/10.1371/journal.pone.0170242>.

Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K, Al-Garadi MA. Classification of forensic autopsy reports through conceptual graph-based document representation model. *J Biomed Inform*. 2018;82:88-105.
<https://doi.org/10.1016/j.jbi.2018.05.008>.

Mujtaba G, Shuib L, Raj RG, Rajandram R, Shaikh K. Prediction of cause of death from forensic autopsy reports using text classification techniques: a comparative study. *J Forensic Leg Med*. 2018;57:41-50.
<https://doi.org/10.1016/j.jflm.2018.03.005>.

Mapundu MT, Kabudula CW, Musenge E, Olago V, Celik T. Performance evaluation of machine learning and computer coded verbal autopsy algorithms for cause of death determination: a comparative analysis of data from rural South Africa. *Front Public Health*. 2022;10:990838.
<https://doi.org/10.3389/fpubh.2022.990838>.

Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Brief Bioinform*. 2018;19(6):1236-46.
<https://doi.org/10.1093/bib/bbx044>.

Prakash JA, Ravi V, Sowmya V, Soman K. Stacked ensemble learning based on deep convolutional neural networks for pediatric pneumonia diagnosis using chest X-ray images. *Neural Comput Appl*. 2023;35:8259-79.
<https://doi.org/10.1007/s00521-022-08038-0>.

Ravi D, Wong C, Deligianni F, Berthelot M, Andreu-Perez J, Lo B, et al. Deep learning for health informatics. *IEEE J Biomed Health Inform*. 2017;21(1):4-21.
<https://doi.org/10.1109/JBHI.2016.2636665>.

Kwak GHJ, Hui P. DeepHealth: deep learning for health informatics. *ACM Trans Comput Healthc*. 2020;1(2):1-22.

Srivastava S, Soman S, Rai A, Srivastava PK. Deep learning for health informatics: recent trends and future directions. In: *Proceedings of the 2017 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*; 2017 Aug 13-16; Udupi, India. Piscataway (NJ): IEEE; 2017. p.1665-70.

Zhao R, Yan R, Chen Z, Mao K, Wang P, Gao RX. Deep learning and its applications to machine health monitoring. *Mech Syst Signal Process*. 2019;115:213-37.
<https://doi.org/10.1016/j.ymssp.2018.05.050>.

Zhang T, Oles FJ. Text categorization based on regularized linear classification methods. *Inf Retr*. 2001;4:5-31.
<https://doi.org/10.1023/A:1011469630455>.

Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification. *arXiv*. 2015. arXiv:1509.01626.

Kowsari K, Brown DE, Heidarysafa M, Meimandi KJ, Gerber MS, Barnes LE. HDLTex: hierarchical deep learning for text classification. In: *Proceedings of the 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*; 2017 Dec 18-21; Cancun, Mexico. Piscataway (NJ): IEEE; 2017. p.364-71.

Mujtaba G, Shuib L, Idris N, Hoo WL, Raj RG, Khowaja K, et al. Clinical text classification research trends: systematic literature review and open issues. *Expert Syst Appl*. 2019;116:494-520.
<https://doi.org/10.1016/j.eswa.2018.09.010>.

Johnson R, Zhang T. Effective use of word order for text categorization with convolutional neural networks. *arXiv*. 2014. arXiv:1412.1058.

Irsoy O, Cardie C. Opinion mining with deep recurrent neural networks. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*; 2014 Oct 25-29; Doha, Qatar. Stroudsburg (PA): Association for Computational Linguistics; 2014. p.720-8.

Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*; 2015 Sep 17-21; Lisbon, Portugal. Stroudsburg (PA): Association for Computational Linguistics; 2015. p.1422-32.

Liu F, Zheng J, Zheng L, Chen C. Combining attention-based bidirectional gated recurrent neural network and two-dimensional convolutional neural network for document-level sentiment classification. *Neurocomputing*. 2020;371:39-50.
<https://doi.org/10.1016/j.neucom.2019.09.042>.

Ghosh M, Sanyal G. Document modeling with hierarchical deep learning approach for sentiment classification. In: *Proceedings of the 2nd International Conference on Digital Signal Processing*; 2018; Tokyo, Japan. New York (NY): ACM; 2018. p.181-5.

Xu J, Chen D, Qiu X, Huang X. Cached long short-term memory neural networks for document-level sentiment classification. *arXiv*. 2016. arXiv:1610.04989.

Jelodar H, Wang Y, Orji R, Huang S. Deep sentiment classification and topic discovery on novel coronavirus or COVID-19 online discussions: NLP using LSTM recurrent neural network approach. *IEEE J Biomed Health Inform*. 2020;24(10):2733-42.
<https://doi.org/10.1109/JBHI.2020.3001216>.

Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*; 2016 Jun 12-17; San Diego, CA. Stroudsburg (PA): Association for Computational Linguistics; 2016. p.1480-9.

Jebblee S, Gomes M, Hirst G. Multi-task learning for interpretable cause of death classification using key phrase prediction. In: *Proceedings of the BioNLP 2018 Workshop*; 2018 Jul; Melbourne, Australia. Stroudsburg (PA): Association for Computational Linguistics; 2018. p.12-7.

Yan Z, Jebblee S, Hirst G. Can character embeddings improve cause-of-death classification for verbal autopsy narratives? In: *Proceedings of the 18th BioNLP Workshop and Shared Task*; 2019; Florence, Italy. Stroudsburg (PA): Association for Computational Linguistics; 2019. p.234-9.

Blanco A, Perez A, Casillas A, Cobos D. Extracting cause of death from verbal autopsy with deep learning interpretable methods. *IEEE J Biomed Health Inform*. 2021;25(4):1315-25.
<https://doi.org/10.1109/JBHI.2020.3011319>.

Kivrak M, Guldogan E, Colak C. Prediction of death status on the course of treatment in SARS-COV-2 patients with deep learning and machine learning methods. *Comput Methods Programs Biomed*. 2021;201:105951.
<https://doi.org/10.1016/j.cmpb.2021.105951>.

Heidarysafa M, Kowsari K, Barnes L, Brown D. Analysis of railway accidents' narratives using deep learning. In: *Proceedings of the 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*; 2018 Dec 17-20; Orlando, FL. Piscataway (NJ): IEEE; 2018. p.1446-53.

El Asnaoui K. Design ensemble deep learning model for pneumonia disease classification. *Int J Multimed Inf Retr.* 2021;10:55-68.
<https://doi.org/10.1007/s40035-021-00264-7>.

Tang S, Wang C, Nie J, Kumar N, Zhang Y, Xiong Z, et al. EDL-COVID: ensemble deep learning for COVID-19 case detection from chest X-ray images. *IEEE Trans Ind Inform.* 2021;17(10):6539-49.
<https://doi.org/10.1109/TII.2021.3066293>.

Ganaie MA, Hu M, Malik AK, Tanveer M, Suganthan PN. Ensemble deep learning: a review. *Eng Appl Artif Intell.* 2022;115:105151.
<https://doi.org/10.1016/j.engappai.2022.105151>.

Baccouche A, Garcia-Zapirain B, Castillo Olea C, Elmaghraby A. Ensemble deep learning models for heart disease classification: a case study from Mexico. *Information.* 2020;11(4):207.
<https://doi.org/10.3390/info11040207>.

Loh HW, Ooi CP, Seoni S, Barua PD, Molinari F, Acharya UR. Application of explainable artificial intelligence for healthcare: a systematic review of the last decade (2011-2022). *Comput Methods Programs Biomed.* 2022;226:107161.
<https://doi.org/10.1016/j.cmpb.2022.107161>.

Zhang Y, Weng Y, Lund J. Applications of explainable artificial intelligence in diagnosis and surgery. *Diagnostics.* 2022;12(10):237.
<https://doi.org/10.3390/diagnostics12020237>.

Javed AR, Khan HU, Alomari MKB, Sarwar MU, Asim M, Almadhor AS, et al. Toward explainable AI-empowered cognitive health assessment. *Front Public Health.* 2023;11:1024195.
<https://doi.org/10.3389/fpubh.2023.1024195>.

Knapič S, Malhi A, Saluja R, Främling K. Explainable artificial intelligence for human decision support system in the medical domain. *Mach Learn Knowl Extr.* 2021;3(3):740-70.
<https://doi.org/10.3390/make3030038>.

Kabudula CW, Tollman S, Mee P, Ngobeni S, Silaule B, Gómez-Olivé FX, et al. Two decades of mortality change in rural northeast South Africa. *Glob Health Action.* 2014;7:25596.
<https://doi.org/10.3402/gha.v7.25596>.

Danso S, Atwell E, Johnson O. A comparative study of machine learning methods for verbal autopsy text classification. *arXiv.* 2014. arXiv:1402.4380.

King G, Lu Y. Verbal autopsy methods with multiple causes of death. *Stat Sci.* 2008;23(1):78-91.
<https://doi.org/10.1214/07-STS247>.

Shah C. A hands-on introduction to data science. Cambridge: Cambridge University Press; 2020.

Naili M, Chaibi AH, Ghezala HHB. Comparative study of word embedding methods in topic segmentation. *Procedia Comput Sci.* 2017;112:340-9.
<https://doi.org/10.1016/j.procs.2017.08.009>.

Alami N, Meknassi M, En-nahnahi N. Enhancing unsupervised neural networks based text summarization with word embedding and ensemble learning. *Expert Syst Appl.* 2019;123:195-211.
<https://doi.org/10.1016/j.eswa.2019.01.037>.

Young T, Hazarika D, Poria S, Cambria E. Recent trends in deep learning based natural language processing. *arXiv.* 2017. arXiv:1708.02709.

Zaki MJ, Meira W Jr. Data mining and machine learning: fundamental concepts and algorithms. Cambridge: Cambridge University Press; 2019.

Leskovec J, Rajaraman A, Ullman JD. Mining of massive datasets. 3rd ed. Cambridge: Cambridge University Press; 2020.

Kirillov A, Schlesinger D, Forkel W, Zelenin A, Zheng S, Torr P, et al. Efficient likelihood learning of a generic CNN-CRF model for semantic segmentation. arXiv. 2015. arXiv:1511.05067.

Malinowski M, Rohrbach M, Fritz M. Ask your neurons: a neural-based approach to answering questions about images. In: Proceedings of the IEEE International Conference on Computer Vision; 2015 Dec 7-13; Santiago, Chile. Piscataway (NJ): IEEE; 2015. p.1-9.

Wolpert DH. Stacked generalization. Neural Netw. 1992;5(2):241-59.
[https://doi.org/10.1016/S0893-6080\(05\)80023-1](https://doi.org/10.1016/S0893-6080(05)80023-1).

Watt J, Borhani R, Katsaggelos AK. Machine learning refined: foundations, algorithms, and applications. 2nd ed. Cambridge: Cambridge University Press; 2020.

Anzanello MJ, Fogliatto FS. Learning curve models and applications: literature review and research directions. Int J Ind Ergon. 2011;41(5):573-83.
<https://doi.org/10.1016/j.ergon.2011.05.001>.

Hoiem D, Gupta T, Li Z, Shlapentokh-Rothman MM. Learning curves for analysis of deep networks. arXiv. 2020. arXiv:2010.11029.