Systematic Health Artificial Intelligence (SHAI) - A pathology based NLP model for improved predictive diagnostics in personalised medicine

Shaista Hussain (timecapp@gmail.com)  
Royal College of Surgeons in Ireland

Research Article

Keywords: Machine Learning, NLP, Text Mining, clinical pathology artificial intelligence, AI diagnostics

Posted Date: September 13th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2049616/v1

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

The Systematic Health Artificial Intelligence (SHAI) model trains on data from medical records and clinical laboratory results to temporally identify disease markers with subsequent pathologies, more efficiently and accurately than is done in the current analog practice. The aim of the SHAI model is to gauge a patient's medical prognostic status based on a conglomerate of data to predict lurking, occult or comorbid pathologies. Newfound associations and predictions would support clinicians in terms of comprehensively visualising a patient's health profile, both in real-time and for the future. Proxy findings would also help to establish personalised references ranges for clinical pathological investigations of body fluids. The SHAI model processes EMR progress text-based notes through a NLP 'Bag of Words' system, which enables the neural network to train in word representation and 'weigh' words of proximity. Using 'forward propagation' of the vectors will allow for output activation from hidden and non-hidden layers of the developing neural network architecture, to then use 'multiclass classification' as the vector contents grow with new data. This manuscript identifies 8 key questions to be addressed by diagnostic ML models and explains SHAI's design as it pertains to maximising human benefit and minimising bias. Despite the automaticity of this laboratory medicine solution, physician end-users remain essential to the diagnostic process and final clinical judgements.

Introduction

The Systematic Health Artificial Intelligence (SHAI) model trains on data from medical records and clinical laboratory results to temporally identify disease markers with subsequent pathologies, more efficiently and accurately than is done in the current analog practice. The Centers of Medicare and Medicaid Services claim that over 6 billion tests are conducted annually from around 170,000 CLIA-certified laboratories in the US alone, highlighting how instrumental pathology results are to the medical decision making process (Hanna and Hanna, 2022). This model is based on a Supervised Learning (SL) - Machine Learning (ML) model using Natural Language Processing (NLP) to code Electronic Medical Record (EMR) and laboratory results.

'Machine learning' was a term that was initially popularised in 1959 by Arthur Samuel as "the field of study that gives computers the ability to learn without being explicitly programmed." In 1997, Tom Mitchell defined ML as "a computer program that is said to learn from experience 'E' with respect to some class of tasks 'T' and performance measure 'P'...where, its performance at tasks in 'T' (as measured by P), improves with experience 'E'." (Liu et al., 2020). The SHAI NLP model's aphorism is then 'E' - the structured and unstructured data models uploaded into the computational system; 'T' - the task of categorising vectors and applying predictive mathematical equations to accrued data points; 'P' - the probability that the model will predict disease outcome.

The evolution of assistive diagnostics systems has been in the making since 1996 when ELIZA, the script-based chatbot released DOCTOR, which could interact with humans, and later in 1977 when MYCIN analysed symptoms of infections to determine causal microbes and suggest drug treatments (Chang et al., 2019). Using universal coding systems for modern data entry, the ramifications of such a model on diagnosis and prognosis could be transformative at levels of research and clinical practice.

AIM

The aim of the SHAI model is to gauge a patient's medical prognostic status based on a conglomerate of data to predict lurking, occult or comorbid pathologies. Artificial Intelligence (AI) applications are increasingly predicting molecular and clinical patterns (Försch et al., 2021); with EMR data, the model would likely accelerate the patient profile, analysis and management process (Velupillai et al. 2018). A purposeful algorithm's utility depends on its 'fit' into clinical work, its ability to improve clinician efficiency, and relieve burdens (Scott, 2021). Therefore, the algorithm must be compatible with existing data systems, adaptable to multiple input styles, and used by trained physicians, to facilitate an ergonomically symbiotic relationship between man and machine. The SHAI model will ideally complement existing workflows, and data output would be intentionally designed for research practicality. Newfound associations and predictions would support clinicians in terms of comprehensively visualising a patient's health profile, both in real-time and for the future. Proxy findings would also help to establish personalised references ranges for clinical pathological investigations of body fluids. Despite the automaticity of this laboratory medicine solution, physician end-users remain essential to the diagnostic process and final clinical judgements.

Model

Structure and Training

Training SHAI requires detailed precision combined with an artistic sense of the grand picture. Both the laboratory and progress notes would amount to big data, and would have to be categorised temporally to differentiate between retrospective and active records.

The SHAI model processes EMR progress text-based notes through a NLP 'Bag of Words' system, which enables the neural network to train in word representation and 'weigh' words of proximity. This involves identifying each and every word as one bit that is turned on in the vector, then the machine translates words into continuous vectors (semantically close words). NLP is most useful for this type of language classification, therefore, the classification must be specific to use the model for predictions in the future based on continuously updated variables, eg. Liu et al. (2017)'s lab test extraction system for $HbA1c$ used an open source NLP to extract lab test variables, numeric values as well as temporal information from both unstructured and structured data. They found that the system demonstrated high precision and recall in $HbA1c$ and glucose for diabetes patients. With the surplus of laboratory investigations conducted on patients who receive multiple orders from a multitude of specialists, SHAI would use Symbolic Information Extraction (SIE) for analyses of clinical pathology results, and Bayesian classifiers to analyse free text. Experimental studies have shown that classifiers exhibit good accuracy when used to analyse sentences pertaining to the sample location, disease presence and status of the illnesses (Moscatelli et al., 2018). The SIE method uses syntactic and semantic linguistic information methods detect diagnostic laboratory values from unstructured text (specimens, analytes, units of measures, detection limits, genes, proteins, pathogens), which Kang and Kayaalp (2013) found to have higher recall performance than other methods with statistically significant
SHAI’s SL predictive process uses multivariate linear regression in data training. The goal is to learn a function $h(x) \rightarrow y$ (predictive likelihood), based on $X$ variables (e.g., signs, symptoms, investigation results), leading to $y$ output (i.e., disease), such that the model’s function is a good predictor of disease ($Y$). Using multivariate regression, the continuous input of data would allow for the addition of new variables during the algorithm run, based on:

$$h_{\theta}(x) = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \ldots + \theta_n x_n$$

This assumes the Multivariate Normality property. In this model, $h_{\theta}(x)$, or $Y$, is the hypothesis function (dependent variable) that runs the theta parameters as input values; $n$ is the number of features; $x_1, x_2, \ldots, x_n$ are the independent variables; and, $\theta_0, \theta_1, \ldots, \theta_n$ are the weights. For example, $\theta_0$ is the patient’s disease status, $\theta_1$ is the likelihood of disease based on age, $\theta_2$ as the disease likelihood by genetic gender, etc. and $X_1$ will be age in years, $X_2$ the gender genotype, etc. ($X_2 = 1$). If $J$ is a function of the parameter vector $\theta$, then the cost function equation is:

$$J(\theta_0, \theta_1\ldots, \theta_n) = \frac{1}{2m} \sum_{i=1}^{m} \left( h(\theta, x(i)) - y(i) \right)^2$$

With respect to annotation, SHAI would train the NLP model using large quantities of annotated data and additional hidden layers. The unstructured data are the clinical notes and the structured data are the lab results, and annotation would be needed to be input into the central database for the system to work. 

Doan et al. (2016) did just that for a NLP for Kawasaki Disease (KD) detection, and annotated 22 Emergency Department notes from confirmed KD patients, to identify KD signs from the clinical text and identified high-suspicion patients with a sensitivity of 93.6% and specificity of 77.5% compared to manual clinician review (Doan et al. 2016). EMRs collect personal demographic data such as age, race, gender, social and clinical histories, and mathematically apply these as independent factors to characterise disease features (Cui & Zhang, 2021). Deep layers would be used for monitoring and correlating the lab results to key words from the medical notes e.g. determine level of renal function, stages of anaemia, autoimmunity inflammatory markers, etc. e.g. determine if fever is present or absent, associated with exposure or non-exposure to pathogens. After laboratory information is extracted, it can be fed into the Logical Observation Identifiers Names and Codes (LOINC) universal knowledge base, supported by the ICD-10.

SHAI’s model ‘features’ are the presence of posted words (individual or contextual combinations) that classify a diagnostic strata. ‘Feature vectors’ recognize categorical words (e.g., breath, renal function, etc), while ‘nodes’ are the diagnoses (infarction, acute renal disease, etc). Concretely, the ‘entity linking’ would group “close context” terms from the ‘bag of words’ to define the various clinical components of diagnoses. The ‘entity annotation’ would involve labelling any unstructured sentences from the notes. A LingPipe platform for text/word recognition and topic classification with spelling correction features (Kang et al. 2013) would benefit this case. SHAI ‘name entity recognition’ will involve naming entities in text based on predefined categories (e.g. renal, immune, London, Djibouti) and adding semantic knowledge from regions to account for ‘distributional shift’ - ultimately to help the model to identify and understand subject and context. Sufficient analogue and electronic medical records would suffice as training data, and applied intra-user models may augment personalised health monitoring, given that they require large quantities of longitudinal data for individual users (Velupillai et al. 2018).

Using ‘forward propagation’ of the vectors will allow for output activation from hidden and non-hidden layers of the developing neural network architecture, to then use ‘multiclass classification’ as the vector contents grow with new data. Using gradient descent, the predictive algorithm will seek avenues towards the local minimum for the (multivariate regression) cost function through iterative updates to theta values. $\theta_{(j)}$ will be repeated until convergence is reached, as it is also simultaneously updates for $J = 0,1\ldots n$.

**Big Data Strategies**

Vectorization of the model’s parameters, the instance features and the target values will also help to reduce the burden of the large datasets in this model. Assuming an 0th feature of value 1, the model’s parameter and instance features become vectorized through transposition making the multivariate regression:

$$h_{\theta}(\alpha) = (\alpha^{T}) \cdot \theta$$

The gradient vector can then be computed in a single step. Furthermore, given the large datasets, SHAI employs a ‘Stochastic Gradient Descent’ to scale the data for algorithmic computation. Additionally, feature scaling will allow the algorithm to converge at a faster pace. The SHAI model additionally employs non-linear boundaries because of complex classifiers, and executes a Support Vector Machine (SVM) algorithm to allow for temporal learning and predictive power. For example, the InfoBot at the National Library of Medicine identifies the components of a well-formed clinical question from clinical notes, then invokes a question answering module that extracts answers to the question about the best care plan for a patient (Demner-Fushman, 2009). Over/under-fitting of data would be managed by ‘Regularisation’ to keep all features but reduce the magnitude of $\theta_{(j)}$. After the first level of data training, end users would be feeding patient results - in real time - into the evolving SHAI, which would continue to train the machine, propagating ‘Retraining’.

**Evaluation**

The SHAI model should be deployed only after a vast array of critical evaluations are conducted, assessed and addressed. The first issue to be determined is whether SHAI’s algorithm improves clinical diagnostics. This can be gauged by the model’s relevance to diagnostics, value input, as well as the accuracy and efficiency of the algorithm. Accurate detection of fluctuations in lab results relating to clinical symptoms may improve clinical bias (missed diagnoses), and augment research associations of markers and comorbidities that have been previously discounted. Huang et al. (2020) used violin plot and paired t-tests to compare overall performance across time series models and proved the feasibility of AI-assisted early detection systems for Acute Kidney Injury predictions (using AdaBoost ML). Segal et al. (2020) used a Kidney Failure Risk Equation consisting of variables: age, sex, urine albumin-to-creatinine ratio, and eGFR in individuals with CKD; to predict the risk of ESRD, and the need for dialysis or a kidney transplant within 2–5 years. Their study showed higher predictive values (c-statistics 0.93, sensitivity 0.715, specificity 0.958) than human counterparts alone. The poor prognosis of renal disease diagnosis makes the AI attempt at
earlier diagnosis an imperative path for us to follow. More concrete associations between symptoms and endogenous markers increases the likelihood that early intervention steps can positively influence disease outcome - the goal of preventative and personalised medicine.

This leads to the second evaluation question: Does the model identify new risks and biomarkers? The benefit to laboratory testing can be gauged by progress and value rubrics as AI improves laboratory testing in two ways. Firstly, AI accelerates the testing process, by analysing large quantities of data points much faster than the trained human eye; and secondly, AI has the power to learn as it works, so as to recognize sample patterns more quickly and accurately over time. Because machine learning is not rule-based, but rather, factors in correlations and permutations, it retains the flexibility to allow for various interpretations. So, it is likely that this model will identify risk factors or new association with diseases.

A third question pertains to the global accessibility of the data: Can the algorithm be trained with global data and shared with a global audience? Global imbalances in access to technology is a limiting factor that requires resourceful solutions to optimise common data sharing. Both uploading training data and retrieving output are impeded in environments with limited internet or technology access. With regards to multiple language inputs, automatic text translation from different languages is difficult but attention-based recurrent neural networks (RNN) have shown to successfully attain English-German bilingual evaluation understudy scores (Chang et al., 2019), which will likely improve as models develop. A future permutation of the attention-RNN could be incorporated into a future version of SHAI by using the RNN's ability to maintain hidden states in recurrent layers, which are essentially a summary of all previous input elements. Scientists must “manage change” by adapting to hardware and software needed to manage big data, at a global level.

A fourth consideration in approving a model such as this is whether or not the model is scalable to evolve. Segal et al. (2020) analysed 10 million records from an insurance database using a feature embedding Word2Vec algorithm method to gather temporal variables data in predicting kidney injury: diagnosis, procedures, and medications; achieving a PPV of 0.71. With new discoveries alongside new technologies, the wealth of haematological and chemistry and microbiological data is growing rapidly (per patient) and a sturdy foundational design can sustain an evolving model to adapt to new discoveries. Therefore, scalability is related to model performance evaluation. And model performance valuation will include measures of the end-user's trust in the reliability of the model, by comparing results of different models using statistical expressions. Logistic regression can be used for a baseline comparison (of model vs human output) to evaluate output performance by: (1) area under the ROC curve; (2) Precision or positive predictive value; (3) Recall; (4) F1 score; (5) negative predictive value, and; (6) specificity. ML model performance can be precisely evaluated with great success, for example Gradient Boosting Trees combine several decision trees in a sequence-dependent series, allowing the model to gradually fit so the prediction loss function is minimised using gradient descent (Segal et al. 2020). ML models have the potential to identify patients with a high risk of readmission before readmission takes place and the potential to suggest the most appropriate action to take to reduce the risk of readmission. The risk stratification essentially creates two groups: high-risk and lower-risk, and this is then complimented by suggested therapeutic interventions. To holistically consider the benefits (and risks), the proposed model would use easily accessible clinical data to perform patient risk stratification from large populations, and assign outputs with L1 Regularised Logistic Regression to optimise predictive performance and choose the most predictive set of features of interest. SHAI would employ a Receiver Operator Curve to measure the Area Under the Curve (focussing on low false positives, and high true positive rates, in predicting disease onset in patients, to accurately arrive at a higher positive predictive value rate).

Patient Safety

There are a number of questions to consider in evaluating a diagnostic model for patient safety. The primary question is: Does the system safeguard patients? Or, put another way: Could the algorithm cause harm to people? This can be gauged by safety measures, security protocol, fidelity and confidentiality management practices.

Personalised data, including genomics and disease risk, are sensitive data that must be subject to ethical protocols serving the best interest of the population. AI database breaches would reveal intimacies that could negatively impact patients if censorship and security were not embedded in the project. Adverse outcomes could occur via cyber attacks which could corrupt datasets or software (Scott, 2021); data breaches could subject patients to personalised threats and violated privacy; and it is also possible that medical algorithms could fail to recognize false or biased output - again highlighting the imperative role of the physician to make final judgements. Data sharing is a sensitive issue of context and privilege, who when and how patient data is shared and protected. A vast array of patient data will be input into the system, and benefactors of the algorithm should ultimately be patients. While end-users are assisted with patient plans, patients must retain confidence in the decision making processes, and a challenge is to consider the consequences of high-level sharing. Another factor to evaluate is whether the model will detect truly-at-risk populations accurately. For example, if a poor prognosis is predicted based on inherent risk factors, a patient’s psychological grief response is yet unpredictable. This also speaks to the risk of physician end-user access to multiple levels of patient data, and the security risk that is posed. Multiple stigmas inhibit many from wanting personal health information shared to prevent potentially harmful information leaks that could affect employment, or insurance. Other ethical issues include accountability, risks of potentially erroneous ML outputs, biases and malicious uses of ML. Therefore, a platform developer must be transparent with how data is secured and the benefits of the AI application should outweigh the risks.

Predictive models should minimise the introduction of bias and unfairness into the healthcare provision platform by addressing equity and bias by considering ethical justice, equality, distributional shift and delicate database design. The NLP design must compensate for unconsciously introduced human bias, for example, a patient on medication will have subsequent changes in lab results that are not representative of the disease progression. The model must contextually consider the collected data to specify associations and reduce bias. Another example of unforced bias is training data retrieved from only a subset population representing a minority, specific demographic, making future predictions biased (distributional shift). Identified risks may also subject patients to unfair insurance coverage plans, and this carries tremendous risk to what are already fragile healthcare infrastructures.

Risk Management
The algorithm's foundation must be simple enough to allow for backtracking in case of errors and must be the right fit for the desired short and long term outcomes. In this case, a hybrid 'symbolic and learning language' is suggested to account for the clinical notes and lab results - requiring text and numerical classifications. A patient's personal life features should be accounted for so that the best clinical decision considers life factors in the decision tree. Additionally, the algorithm should consider co-factors that affect detected features, eg... new onset medication or comorbidities, or new epigenetic lifestyle modifiers (eg, smoking, running, etc.). With respect to the data, SHAI would use high-quantity globally-acquired training data, and automate results reporting.

Given the risks, the SHAI model proposes to mitigate risk to patients and stakeholders, through perpetual risk evaluation and respective management. Firstly, the Data Protection Impact Assessment (required when personal data is used by an AI), and could enable a processor-controller dynamic, lined with firewalls and security verification steps. The machine processes and the final outcome would be judged by the physician/controller. The SHAI model requires investment in operations security for all facets of communication systems, to ensure the integrity of data transmission. This process would be designed with legal consultation, especially to define the roles and responsibilities of operators and the machines, in explicit detail at the project onset; to include consequences of non-compliance to responsibilities. SHAI would also undergo regular performance evaluation by end-users, benefactors and stakeholders to ensure that outputs align with operational requirements and key performance indicator metrics. The list of stakeholders includes pharmaceutical companies looking for targeted therapeutics; health insurers for claim verifications; healthcare administrators who work to streamline patient processes; policy makers who make decisions based on research and to reduce risk and improve human control; and ethics boards to include legal counsellors. A supervisory role for human and machine would involve enabling quality checkpoints at each step of the workflow and algorithm, to allow time to implement required resources and make corrections - this requires teamwork and efficient communication amongst all involved parties. An additional supervisory role will be assigned to a ML engineer for anomaly detection within the algorithm.

'Explainability' of the model's decision tree, or the map of what the machine is doing to reach an outcome, could beneficially include familiar interpretation methods including statistical analyses, visualisation graphs and stakeholders’ success phenotype markers (key performance indicators). Common 'explainability' communication methods include weight histograms, saliency and occlusion maps, class or activation maximisation, Local Interpretable Model-Agnostic Explanations, Partial Dependency Plots, Individual Conditional Expectation, Shapley Values, anchors, counterfactuals, and model distillations (Steward, 2020). Yuan et al. (2020) showed that ML models improve anaemia outcomes in hemodialysis patients, minimising prescription of erythropoietin agents, with the potential to reduce treatment cost - in this case, the investors and pharmaceutical companies could yield evidence to support the investment. Clearly communicated definitions of expectations from respective stakeholders will allow model designers to tailor data outputs.

**Conclusion**

For the Systematic Health Artificial Intelligence model to be useful to end-users, a pilot test should be conducted using medium-size regional data for a locale. In the beta run, real-time predictions would allow for troubleshooting, version permutations and feasibility evaluations for the mathematical model, the code and the end-user interface. While the model's predictions are made in large part by the mathematical formulae, the operation also relies heavily on the type of data being input and the structure of databases receiving the data. Therefore, physician end-users should have a solid understanding of the data feeding system into ML models as well as the ability to discern output contextually. Man-Machine communication must include elements of interpretability and trust in information exchange. Steward (2020) suggests evaluating model interpretability with a quantifiable proxy, so perhaps alert notifications flagging increasing likelihood of disease acquisition would hasten the diagnostic process. The communication would also require a user-friendly, graphical user interface - which, if smartphones are indicators, can even foster machine dependency.

Later, when the 'Moore's Law' principle is effected, albeit exaggerated with AI technology, the system would be cheaper to operate and could be applied to animal health too. More intense testing is recommended as the system evolves from simple diagnostic applications to automated applications (Scott, 2021) that can determine a wider array of disease markers. The value-added benefit of this clinical lab and notes system is the algorithm's ability to predict and diagnose faster and more specifically than a human system. The proposed SHAI model uses an NLP structured model to consort heterogeneous data sources from a patient's clinical journey, using both structured and unstructured data aggregation methods to perform diagnostic and prognostic predictions in personalised medicine.

**Practice Points**

There are a number of key questions to be considered in the evaluation of a proposed ML model to determine the likelihood of success in healthcare operations. These applied to the SHAI model, include:

1) Does the algorithm improve clinical diagnostics?
   a) This can be gauged by the model's relevance, value, accuracy and efficiency.

2) Does the model identify new risks and biomarkers?
   a) Laboratory medicine benefits can be gauged by progress and value rubrics.

3) Does the model detect truly at-risk populations accurately?
   a) This can be gauged by justice, equality, distributional shift and the handling of training data.

4) Does the system safeguard patients - could the algorithm cause harm to people?
a) Whether direct or indirect, this can be gauged by safety measures, security protocols, fidelity and confidentiality practices.

5) How are risks to stakeholders and patients being moderated?

a) This risk evaluation is a priority for management to delineate the risks and outcomes to guide the development of the next phases of the model

6) Does the model evolve?

a) The scalability of a model is largely dependent on the foundation and the quality and quantity training data that is appropriately used.

7) Is the data globally accessible?

a) Accessibility is both for data input and predictive outputs, and this can be gauged by the practicality of the model and the internal validity of the algorithm, as well as outcome predictions, distributional shift and change management strategies.

8) Does the model effectively communicate findings with the end-users?

a) This domain speaks to interpretability and trust, and the graphical user interface (GUI) and design leading to a fulfilling user experience (UX) serve as meeting points for machine and human user.

Declarations

Competing interests

I declare no competing interests.

Funding

I declare no funding received or associated with this paper.

Ethical Approval

No ethics approval is required for this review.

Consent to Participate

Not applicable

Consent to Publish

I consent to have this published.

References


