Leveraging on the Mining of Locally Generated Huge Databases for AI Driven Healthcare in Emerging Economies

Michael Sunny Osigbemeh
Dept. of Electrical & Electronic Engineering
Alex Ekueme Federal University
Ebonyi, Nigeria
eosigbemeh@yahoo.com

Alpheus Ahaku
Dept. of Chemical Engineering
Alex Ekueme Federal University
Ebonyi, Nigeria

John Ogbemhe
Dept. of Systems Engineering
University of Lagos
Akoka, Nigeria.

***Abstract*—The enormous possibilities of Big data and mining its cache was used to identify patterns in mostly non formal diagnosis and medications in resource constrained African communities that are largely classified as emerging or developing economies. Dependence on Neurofuzzy analysis have resulted in the acquisition of largely public available though not formal medical data for further fuzzineural analysis after acquired data has been preprocessed for vagueness, consistency and relevance. These acquired data forms the inputs to the designed fuzzineural system for further processing to obtain expert advice on presented symptom by current patient on diagnosis. The resulting expert system while leveraging on connected dots in the pre generated hypothetical database makes an informed decision on presented and inputted symptoms after attaining optimal iterations through training and validation. These results can then serve as extra support or pre validation action for healthcare givers.**

***Keywords*—Expert System, Neurofuzzy, Big data, Healthcare, data mining, machine learning, ANFIS**

I. INTRODUCTION

Emerging trend of social activities especially online and electronic enable contacts and interactions has continued to point today's youth towards heavy reliability on less traditional learning process. This trend has necessitated dependence on cheap easily sourced materials, information and data from mostly online means for treatment and diagnosis of current and longtime ailments. Online data and information which sometimes also contains prescription doses are mostly unverified and unratified diagnosis leading to self-prompted diagnosis and medications.

A review of the trend of dependence on informal sourced diagnosis and suggestive medications especially through online sources have indicated a very high trend for treatment of various health issues including death threatening situations especially among the resource constrained population such as in Nigeria and other sub Saharan African countries battling with high poverty rates among citizens.

Further debilitating the situation is the apparent non commitment in educating the populace on the dangers of self-medications by the governmental agencies and the adverse and glaring departing of skill medical practitioners to foreign economies for better renumeration leading to the popular 'japa' syndrome.

Surprising, these informal sourced diagnosis and medications have been responsible for significant numbers of healthcare among the elderly and youth population with largely unavailable success rates and also with undocumented hazardous occurrences and outcomes.

A deeper analysis of the situation suggests that the dependence on this style of treatment by emerging economies are supported by data sources and practices passed down by heuristic means from generation to generation that leverage on traditional structures, trial and error experiences, spiritism, etc. leading to known pairs and combinations of drugs, foods, drinks, etc. for treatment and longevity. These input-output pairs have been directly linked with mostly success rates when compared with hazardous or life-threatening consequences thereby necessitating research into this emerging area or way of life.

The disease or ailment rate, propensity and gravity are however nonlinear and are influence by the population in different modes based on underlying issues, living status, income, weather, emotional factors, etc. These various participatory parameters lead to great complexities and non linearity in the analysis and administration of informally sourced diagnosis and medications. Approximations of nonlinear problems which are common in nature and real-life situations have been analyzed by [Witten *et al*, (2011); Tzafestas *et al*, (2002)] who suggested using mathematical approximations such as fuzzy logic processing real world data which are highly non linear. Computer aided analysis and the role of mathematical operations in sieving through huge data sets have been elaborated by [Zhou *et al*, 2015, Han *et al,* 2012; Nisbet *et al,* (2009)] to obtain hidden knowledge and harvest connected dots.

Given the above premise it is important for the populace of resource constrained developing economies such as Nigeria, through NGOs, private and governmental institutions to begin to document diagnosed ailments and treatment pairs along with their predisposing factors in a huge structured database that allows updates from several unconventional sourcing but which a data management and sorting algorithm in place to prone out mis use or misinformation. This massively generated data source would then be mined and used to provide healthcare support to vulnerable people in the event of demand for healthcare in an extreme unavailability of medical personnel and healthcare givers by leveraging on the principles of data mining and machine learning to 'suggest' a real-time diagnosis of occurring ailment to the current patient.

II. METHODOLOGY

A. Artificial Neural Network and Fuzzy Inference System

Preprocessed data that are largely from vetted public domain are fed into artificial Neural Network and fuzzy inference system (ANFIS) to obtain validation of processed symptoms of real-time patient. The fuzzy inference system (FIS) is responsible for aggregating spurious nuance from largely vague and ambiguous predicates associated with collation of data of the form this work intends to acquire and process. Data acquisition sources such as administration of questionnaires, direct data sourcing, data from field trips and meetings, medical records, etc. are preprocessed for ambiguity, inconsistences, vagueness, relevance, incompleteness, etc. to form a fuzzified and normalized input pairs of symptoms for further processing.

The artificial neural network (ANN) would then be fed with the pre-normalized and fuzzified inputs now domiciled on the database for further analysis and diagnosis. The ANN leverages on the huge data plausibility of connected dots in the specific database from collated several millions of entries nationwide by setting aside a training set of 50%, a testing set of 25% and final 25% set for validation [Larose (2006); Osigbemeh *et al* (2017)] of diagnostic outcome or result of iterations.

The various sets are inputted into the resulting ANFIS system and successively iterated to attain winning epochs for each successive set fed to the ANFIS topology during processing.. Also, the type of convergence sought after in terms of speed and accuracy of iterations and subsequent diagnosis could be obtained from guidelines set by [Osigbemeh et al, 2024; Rudin & Wagstaff (2013); Suzuki et al, (2005)], that set various parameters such introduction of learning rate term or/and momentum term in the formula for solving gradient descent optimization technique for back propagation algorithm learning.

A general architectural topology for iterations and network convergence for processing of the pre generated database of symptoms and diagnosis pairs is shown in fig. 1. Cleaned up, normalized and verified diagnosis of ailments serves as inputs x_i for *n* numbers of symptoms used in diagnosing or identifying current ailment. With a single hidden unit layer for fast convergence since database contained presorted data, ANN processing converges to a single output indicative of necessary action. This is particularly useful for timebound contagious cases making first contact with healthcare institution.

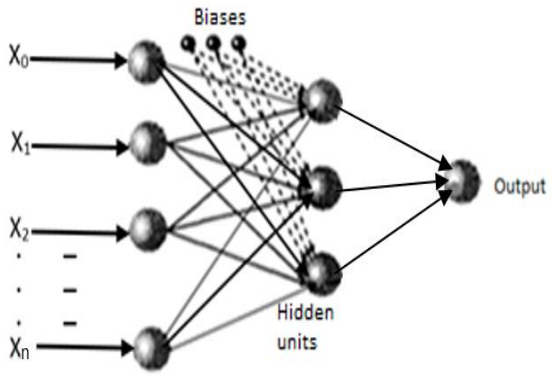

Fig. 1. Topology of generalized ANN

B. ANN and Network's Architecture

To obtain a winning epoch and therefore a possible diagnosis, for the current input of symptoms, the ANN architecture must be chosen with factors of speed, accuracy and validation in mind. A possible diagnosis can be obtained from the mined database mathematically by the form:

$$f(z) = \sum_{i=0}^{n} xi \quad (1)$$

Where f(z) represents the desired diagnosis and x*i* represents the participating or contributory symptoms. As shown in fig. 1, $xi = x_0$ is the first contributory input symptom into the ANN for that current ailment. This was closely followed by x_1 and so on. A gradient descent optimization scheme that utilizes the back-propagation algorithm though slower in attaining convergence than the fast in convergence of the Levenberg-Marguart algorithm (LMA) could be chosen by the resulting expert system depending on the nature of presented ailment, time constraint, available resources, etc; however, depending on the available time for the diagnosis the LMA will be the method of choice by system's default to obtain faster convergence for a time conscious assessment while pending BPA would be used to ratify LMA processed results. This is particularly important where patients ailment is timebound or contagious nature and requiring immediate diagnosis.

C. System Architecture and Data Retrieval Process

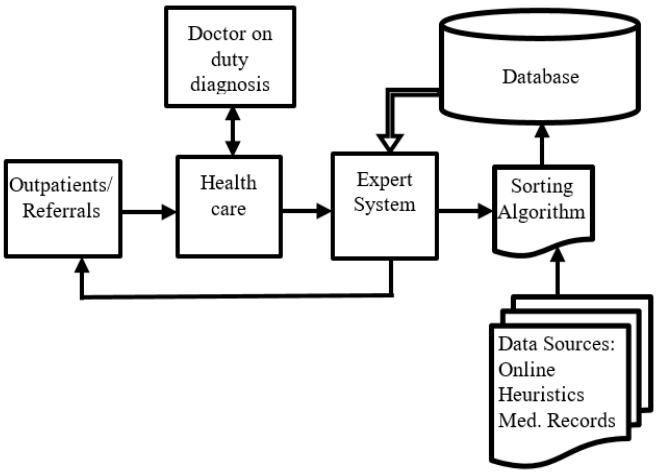

Fig. 2. Block diagram of overall process

Diagnosis begins when an outpatient or patient on referral cannot be attended to by the resident doctor due to unavailabilityor scarcity as established above, the healthcare then activates the medical expert system were presented symptoms are fed and the expert system tries to iterate through its database to converge to an optimal diagnosis for attending to the patient under current diagnosis. Considering the expected huge database of symptoms/diagnosis pairs, the sorting algorithm ensures proper storage or domestication of data and appropriate retrieval of data in real-time. As shown in fig. 2, the database is still being updated by online sources, specific diagnosis based on heuristics, available medical records from previous healthcare with similar symptoms/diagnosis. The sorting algorithm also handles the data quality of the data under reposition to ensure conformity, cultural barriers, popular beliefs, existing physiological differences and communal factors that may directly influence diagnosis including previous medication by patent are properly documented into database. Known outbreaks of diseases in the human populations with epidemic propulsions have been observed to occur by healthcare giver's error, negligence and lack of attention to patients presented symptoms' details.

III. RESULTS AND DISCUSSION

An ubiquitous expert system of this form must be able to make inform and professional diagnosis in line with the known characteristics of expert systems by sieving through every detail in inputted systems including information on zoonotic behaviors and encounters of patient in the case of emerging contagious nature of diseases based on processed information when compared from the database. These results could be revalidated by medical personnel if (or when) available physically or online.

The resulting expert systems has been pre-equipped to process confidence rates or levels based on presented symptoms to ensure that current diagnosis is based on overwhelming validations from training dataset, testing and validation datasets from the database. The graphical user interface (GUI) which serves as the window into the expert system is allowed to display the confidence level of ANFIS processing to the healthcare consultant based on inputted symptoms. These confidence levels suggest to the consultant the severity levels of current diagnosis including suggestions on safety in a contagious session which is basically initialization of quarantinement procedures.

A. Challenges with data acquisition systems

Data mining systems and machine learning methods approach excellence and robustness in coherence, relevance and consistency in acquired data for processing. The developed ANN and resulting ANFIS gave appreciable results of diagnosis of ailments in real-time when data was of integrity and unambiguous nature allowing users to visually detect its processing and concluded results through the GUI.

The arduous task of gathering or collection of data pairs of ailments with presented symptoms and medications or solutions from formal and informal sources to attain the characteristics of big data in a database is very daunting. This is due to the need for a robust sorting system, sources of input validation, identification of bots and other malwares trying to overwhelm database storage system, etc.

B. Robustness of ANN analysis and results

Processed results from BPA are less susceptible to spuriousness when compared to LMA which experience tradeoffs due to extreme speed of operation and convergence on solving for the Jacobean of generated matrixes by the inputted symptoms and iterated database pairs. However, both models were used to determine results of current diagnosis based on several factors such as availability of computer resources, contagious nature or symptomatic nature presented, need for validation and assurance, etc. bearing in mind the fact that hidden behind big data are enormous information and connected patterns. By leveraging on the speed of LMA processing, presented symptoms can be quickly brought to convergence, results obtained and preliminary actions initiated while awaiting the result of GDO processing.

C. Discussion

The resulting medical expert system allows for keyboard or other input unit support of real-time occurring symptoms that are visible or being experienced by current patient. These inputs are displayed on the GUI just like any expert system and execution is allowed to commence, After aggregation of inputted data, the expert system is allowed to analyze the inputs after pre-fuzzification, normalization and iterations to attain winning or optimal epoch by computing equation (1) for all inputted symptoms and iterating with the training data that consists of 50% of big database, then testing with 25% and validating with 25% of big database. Table 1 shows a sample of the generalized syntax for linguistic data input to the FIS part of the medical expert and how it determines the fuzzification and normalization values for input into the ANN for further processing to generate the optimal solution and subsequent suggestion on diagnosis and medication.

In Table 1, responses to ANFIS medical expert shows four symptoms successfully acquire from a hypothetical patient with possibility of increasing the number of symptoms if necessary. Also, weighted fuzzified values are mathematically generated based on predefined relationship and connected considerations based on the principles of fuzzy logic and fuzzy sets. More symptoms input will guarantee a more efficient representations of occurring or presented symptoms for further processing by ANN section of the medical expert especially for zoonotic considerations and epidermic periods.

TABLE I. SYNTAX OF SYMPTOMS PAIRS FOR EXPERT SYSTEM

SN	**ANFIS Medical Expert**			
	Presented symptoms	***Duration***	***Degree***	***Fuzzified value***
1	Body temperature	3 days	Very High	0.7
2	Fever	2 weeks	Absent	0.3
3	Any contact with sick patient	3 hours	Closed contact	0.9
4	Has symptoms occurred before	N	-	0.6
. n	. n	. n	. n	. n

D. Output Flow of Diagnosis

It must be reemphasized that current diagnosing plausibility from the medical expert must be accepted with a pinch of salt as results from ANFIS processing would only be considered in the event of a complete absence of medical practitioners and only as a suggestive measure due to the non-crisp or dynamic nature of diseases and ailments as presented by patients or on the human population. An error in diagnosis value (EIDV) which is a value or percentage that indicates the level of confidence of the resulting expert system's session as discussed above also help to improve the expert system consultation. The EIDV is an attempt to fuzzify the resulting expert system's processing to attain the nuance or ambiguity in human related responses to interaction. This value if very high for index cases of presented symptoms can be used to advice the healthcare givers to quarantine a patient on current diagnosis if the presented symptoms are suggestive of an epidemic outbreak or of a contagion history with zoonotic tendency.

E. The Medical Expert Graphical User Interface

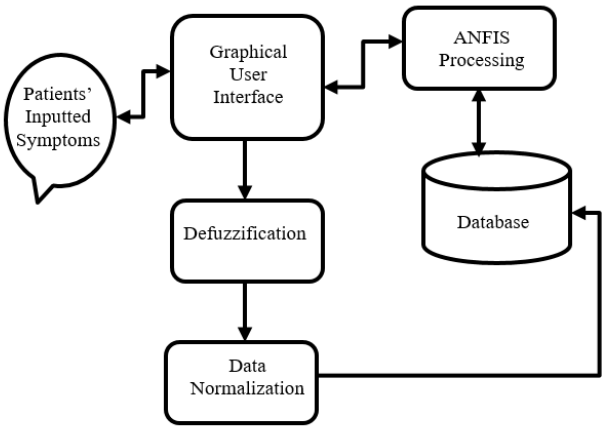

Fig. 3. ANFIS Processing Architecture

Embedded control instructions domiciled in the ANFIS architecture allow for real-time modelling of presented symptoms as inputted by patient or attending healthcare giver. Sessions inputs form basis of convergence criteria for iterations and convergence by the artificial neural network as indicated in fig.1.

Sessions are recorded or stored as new database content for a more robust database management system. Data normalization and defuzzification help to make the inputs admissible to ANN and amenable for further processing. Proning of unconnected dots in acquired data or detection of usuable repository helped to ensure the integrity of data capture and informed decisions by the Anfis medical expert.

IV. CONCLUSION

With the increasing capacity of memory storage systems in terabytes of data storage in hard disk drives and availability of cloud-based storage services coupled with a symbiotic increase in knowledge-based systems of machine learning, big data opportunities and mining, an expert system capable of analyzing huge data from largely heuristic sources of informal sourced diagnosis, occurring symptoms and medications from public domain is presented. The resulting ANFIS medical expert system is capable of analyzing pre-fuzzified linguistic nuance data domiciled in a pre-collated database as inputs to its ANN architectural system for attaining a winning or optimal convergence that indicates a diagnosis of presented symptoms through an iterative scheme. The results are useful in resource constrained emerging economies with challenges of healthcare availability, cost of healthcare and migration of healthcare givers. Thus, this expert system aims to leverage on publicly available pairs of diagnosed ailments and verified treatment options to suggest an informed diagnosis and medication based on machine learning of connected dots or patterns in an aprori collated database. With plausibility for identifying index of contagious disease cases to allow for early quanrantiment of presenting patient by the error in dignosis value computed by the medical expert and points to the confidence level of diagnosed symptoms.

REFERENCES

[1] Witten, I.H; Frank, E. & Hall, M.A. (2011). Data Mining: Practical Machine Learning Tools and Techniques. 3rd Ed. Morgan Kaufmann Publishers-Elsevier Inc. Burlington, USA.

[2] Zhou, Z., Chawla, N.V., Jin, Y., & Williams, G.J. (2015). Big Data Opportunities and Challenges: Discussions from Data Analytics Perspectives. *IEEE Computational Intelligence Magazine*. 9(4), 62 – 73.

[3] Teorey, S., Lightstone, S., Nadeau, T. & Jagadish, H.V. (2011). Database Modeling and Design: Logical Design. 5th Ed. Morgan Kaufmann Publishers-Elsevier Inc. Burlington, USA.

[4] Tzafestas, S.G., Skoundrianos, E.N. & Rigatos, G.G. (2002). Nonlinear Neural Control of Discrete Time Systems using Local Model Neural Network. *Intl. J. Knowledge-Based Intelligent Engineering Systems*, 4(2), 130 – 140.

[5] Osigbemeh, M.S., Ohaneme, C.O. & Inyiama, H.C. (2017). An Algorithm for Characterizing Prefuzzified Linguistic Nuance using Neural Network. *Springer International Journal of Speech Technology* 20(2), 355 – 362. DOI.10.1007/s10772-017-9413-5.

[6] Han, J., Kamber, M. & Pei, J. (2012) Data mining: concepts and techniques. 3rd Ed. Morgan Kaufmann Publishers-Elsevier Inc. Waltham, USA.

[7] Larose, D.T. (2006). *Data Mining Methods and Models* (pp 204 – 239). New Jersey, U.S.A: John Wiley & Sons, Inc.

[8] Nisbet, R., Elder, J. & Miner, G. (2009). Handbook of statistical analysis and data mining applications. Academic press-Elsevier Inc. Burlington, USA.

[9] Osigbemeh, M.S., Osuji, C.C., Onyesolu, M.O., Onochie, U.P (2024). "Comparison of the Artificial Neural Network's Approximations Based on the Levenberg-Marquardt Algorithm and the Gradient Descent Optimization Method on Datasets." Artificial Intelligence Evolution, 5(1), 24–38. DOI: https://doi.org/10.37256/aie.5120243781.

[10] Rudin, C. & Wagstaff, K. L. (2013). Machine Learning*, Special Issue on Machine Learning for Science and Society*. Springer. DOI 10.1007/s10994-013-5425-9

[11] Suzuki, K., Shiraishi, J., Abe, H., MacMahon, H. & Doi, K. (2005). False-positive Reduction in Computer-aided Diagnostic Scheme for Detecting Nodules in Chest Radiographs by Means of Massive Training Artificial Neural Network *Academic Radiology*, 12(2), 191 – 201. doi:10.1016/j.acra.2004.11.017.