Received: 5 February 2024 | Revised: 11 March 2024 | Accepted: 16 April 2024 | Published online: 22 April 2024

RESEARCH ARTICLE

Medinformatics
2024, Vol. 00(00) 1–8
DOI: 10.47852/bonviewMEDIN42022583

BON VIEW PUBLISHING

Decoding Medical Diagnosis with Machine Learning Classifiers

Subhagata Chattopadhyay[1,*]

[1]*Independent Researcher, India*

Abstract: Clinical decision support systems (CDSS) are gaining popularity in disease screening and grading in the current era of digital healthcare. This paper attempts to model how a computer learns to grade an infectious disease (ID), e.g., typhoid fever using the machine learning (ML)-based approach mimicking how a novice doctor learns to diagnose a case with the help of senior doctors. To achieve the goal, ten virtual junior clinicians are developed using ten machine learning classifiers (MLC)-based CDSS, which are then trained with "weighted" [0,1] sign symptoms and the corresponding "labeled" grade of synthetic typhoid fever cases ($N = 198$). Weights and labels are assigned by ten senior clinicians providing their rich clinical knowledge base. The performance of each VJC is then measured in terms of their diagnostic accuracy, precision, recall, and F-score. Results show that random forest (RF, i.e., VJC9) and decision tree (DT, i.e., VJC4)-based CDSS can grade with an average of 87% accuracy, which is even higher than human clinicians' accuracy. The reason behind RF and DT's appreciable performance is that clinicians use tree-search-based methods with probabilistic "yes" and "no" logic to learn the disease patterns alike the working principles of DT and RF for diagnosing and grading any ID. Apart from modeling, the paper provides insight into how to select the right machine learning classifier (MLC) algorithm in the field of ID diagnosis. It also throws light on various hardships and challenges with MLC-based CDSS implementations in the real-world scenario.

Keywords: typhoid fever, infectious disease, machine learning classifier, clinical decision support system, virtual doctor, learning algorithms

1. Introduction

Typhoid fever, also called enteric fever, is a systemic infectious disease caused by the bacteria called *Salmonella enterica* serovar *Typhi* with a total annual global incidence is about 11–21 million cases leading to 128,000–161,000 deaths [1]. It is a gram-negative bacterium that enters the small intestinal lymph nodes via food and water, colonizes in the reticular endothelial cells, and from there enters the bloodstream causing toxemia where the bacterial toxin is liberated in the blood stream [2]. Typically, symptoms of typhoid range from prolonged fever, frontal headache, generalized debility and malaise, loss of appetite, nausea, jaundice, abdominal pain, neurological complications, and in severe cases intestinal perforations and death [3]. Preliminary diagnosis of typhoid fever is essentially clinical, i.e., based on interpretations of patients' sign symptoms made by the clinicians [4]. The Widal test is useful for screening when the clinicians strongly suspect typhoid as the possible cause behind the signs and symptoms [5]. However, in practice, Widal is not so dependable due to the occurrence of false-negative and positive results [5]. In this context, it is important to state that with an upper bound of ≥ 200 for "O" (somatic antigen) and ≥ 100 for "H" (flagellar antigen) agglutinin, the Widal test can correctly screen approximately 74% of cases, which means that there is a 26% error in detection [5, 6]. Hence, blood culture remains the primary diagnostic test [7], while bone marrow culture is the gold standard for diagnosing typhoid fever [8]. However, it is time consuming and within the period of giving the sample for the test and receiving the report, patients may rapidly deteriorate if treatment is not initiated. Another important condition is that typhoid symptoms at the early stage mimic other infectious diseases (ID), e.g., rheumatic fever, tuberculosis, malaria, and many viral illnesses such as hepatitis. Such camouflages may lead to errors in the diagnoses costing time and lives. Therefore, not only early screening of typhoid but also detection of its "stage" at the time of the first diagnosis made is important and remains a clinical challenge to the doctors.

Digital health is evolving continuously from telemedicine and analytics to diagnostic decision support systems (DDSS) using machine learning classifiers (MLC), especially for predicting the type and grade of complex systemic illnesses. The objective behind using MLC algorithms is to provide a faster and more accurate screening and grading of a rapidly spreading disease as an assistive tool to clinicians [9]. As ID progress fast within an individual and spread in the community, it mandates a faster diagnosis and start of treatment as early as possible [10]. Here DDSS can play a major role, especially in the remote areas where clinicians are not readily available or a second opinion is sought for diagnosis and referral to a higher center to provide high-end medical facilities. Typhoid fever is a classic example of a systemic infectious disease that often leads to death if not treated at its early stage. Hence, DDSS tools can be proposed to be the "assistive tools" to the healthcare staff (e.g., novice doctors and even nurses) in preventing further worsening the morbidity and mortality in the population at risk [11]. It is also important to mention here that, empirically used antibiotics in erratic doses, i.e., wrong strength and duration, antibiotics salad as poly-pharmacy popularly known as "PolyRx" are the leading causes of multidrug resistant (MDR) typhoid, especially in the countries where antibiotics are available "over the

*Corresponding author: Subhagata Chattopadhyay, Independent Researcher, India. Email: subhagata.chattopadhyay2017@gmail.com

counter", and the number of clinicians is less compared to the size of the population as seen in the developing nations. People are eventually pushed for self-medication as the only choice [12]. Therefore, a timely diagnosis of typhoid by experienced clinicians is crucial for its prevention, and DDSS can be useful to assist. Although vaccines against typhoid fever are available, their efficacy is not 100%, and boosters are required every 2 years and 5 years for injectable and oral vaccines, respectively, which many do not pursue [13]. It is important to note that apart from the MLCs, rule-based [14] and case-based systems [15] are also used to create DDSS or CDSS to streamline the diagnosis and management. The key issues with these types of systems are the (a) inherent hardness within the rules/cases when triggered, (b) too many rules, depending on the clinical scenarios often introduce confusion, and (c) the complex mechanism of rule pruning to reduce the computational complexity and thereby facilitate the faster decision-making process.

Current literature on applied MLC algorithms for typhoid fever diagnosis does not score much, compared to other communicable diseases. However, due to space constraints, some relevant studies on ID are discussed below and readers may note this is not an exhaustive list.

A bioinformatics-based decision support system (BDSS) using *ANOVA* to diagnose malaria, typhoid, and a combination of malaria and typhoid has been found 97% accurate in identifying the diseases [16]. To differentiate between the proposed and orthodox diagnostic systems, *t*-statistics has been used, which shows that the BDSS is different from the orthodox system and more efficient. In another study, the *Reduced Error Pruning Tree* was used to detect typhoid fever in the Nigerian population with 100% and 98.6% accuracies with training and testing data, respectively [17]. The same set of researchers has developed a *GUI-based typhoid grading system* on labeled typhoid data using Visual Basic, and the diagnostic accuracies are above 95% on training and test data [18].

A detailed survey of the old and current literature finds three existing *research gaps*—i) none of the MLCs are trained with the "perceptual or weighted" real-world clinical rule base (CRB) of the SC, (ii) mathematical modeling of doctors' diagnostic process (see Figure 1) using MLCs, which is novel, and (iii) exploiting the advantages of several types of MLCs to identify the most appropriate one or in combination (ensemble models) for ID diagnosis and grading where speed and accuracy are critical to curbing further spread in the community. This paper is an attempt to address these gaps as novel contributions to ML research in diagnostic healthcare.

2. Methodology

The concept of DDSS is gaining importance due to the advent of ML and AI in healthcare to minimize the gap between the shrinking healthcare workforce and rising demand of delivery across the globe. To develop a DDSS, it is pertinent to know the process of how a clinician arrives at a diagnosis. It is to note that the author of this paper is a medical doctor having expertise of applying AI/ML to conceive, design, and implement DDSSes across clinical scenarios and needs. He has conceived this idea since COVID-19 period and started working on it as an independent researcher. He noted a rise in typhoid fever during that epidemic but overshadowed by the overwhelming impact of COVID-19. He also felt due to imposed social distancing and lockdowns when patients cannot interact with the doctors in person for getting physically examined, virtual clinicians powered by embedded MLC algorithms could be useful to the population on a high demand for medical opinion. Based on the novel concept, in this paper, ten virtual junior clinicians (VJCs) are developed with ten standalone MLCs in grading typhoid fever as DDSS to model how a computer can be taught to diagnose and grade an illness using MLCs similar to how a novice doctor learns diagnosing disease with the help of SC. The concept behind the methodology can be seen in Figure 1.

In this figure, the steps of the training and testing process of VJCs are shown for DDSS. To model it, each column-wise "scaled" input sign symptom (i.e., clinical database or CDB) has been represented by a Gaussian distribution (refer to Equation (1)) with the centroid or mean (μ) having the highest weight (1.0) and the standard deviation or spread (σ) is the stretch of the weight until close to "0" seen in Equation (1). Therefore, μ denotes the classic sign symptom (i.e., salient features) of the illness as it possesses the highest value.

$$w_{x_i} = \mathrm{f}(x_i) = \mathrm{a}.\mathrm{e}^{\frac{-(x_i-\mu)x_i{}^2}{2\sigma^2}} \tag{1}$$

In this equation, "a" is the height of the curve's peak, which is 1.0 here and also represents μ; "e" is the Euler's number; *f(x)* is the function of "x_i" where "x" is an integer value; "i" varies from 1 to N; and w_{x_i} is the weighted sign and symptom.

As the sign symptoms are never binary, i.e., "0" and "1" in reality, their weights practically fall between "0" and "1", which is a float value. Now, any sign symptom (x_i) can be mapped based on the distance from the μ within the spread of σ and weight can be

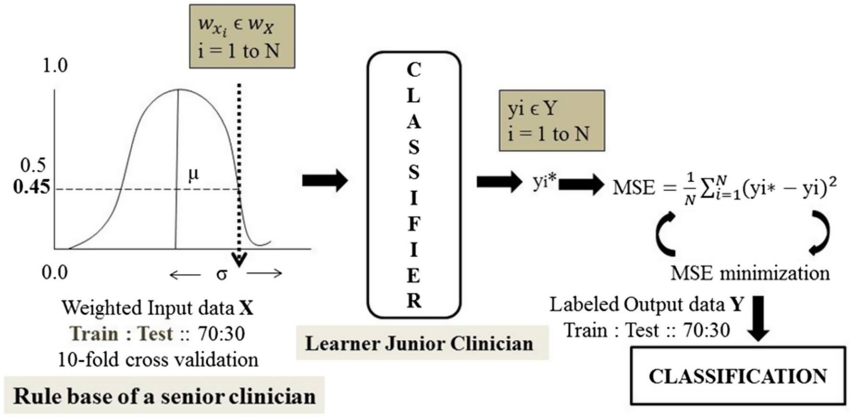

Figure 1. The classification tasks

assigned to it. "x_i", then becomes "w_{x_i}". The paper proposes that this is how an SC assigns weight to each of the signs and symptoms while assessing any ID and also other illnesses. The CRB of an SC is then a set of "w_{x_i}" and its corresponding grade (y_i), which is learned by the VJCs. In other words, the resulting rule can be "for a set of antecedents 'w_{x_i}' ($i = 1$ to N), its respective 'consequence grade (y_i) is labeled as either mild, moderate, or severe'". During the knowledge acquisition by a VJC, the CRB is taught by the SC as "supervised learning", and it helps constitute their CKB, where CKB is a combination of CRB and CDB. In the medical school, the training of the students happens in the similar fashion where SC teaches how to assess individual sign and symptoms to eventually grade an illness.

On this light, each of the ten MLC algorithms, i.e., ten DDSSes that represent the VJCs have been trained by the CKB of the SC in grading typhoid fever, and their performances are compared. 70% of the CKB is used to train the VJCs, while their performance is tested with the remaining 30% of the CKB.

2.1. Data collection

CDB ($N = 198$) had been generated synthetically by a set of experienced general physicians mimicking cases of typhoid fever, i.e., all patients are blood culture and Widal test positive. The CDB has consisted of eleven sign symptoms, such as (i) fever *(FV)*, (ii) headache *(HD)*, (iii) weakness *(WK)*, (iv) muscle pain *(MS)*, (v) sweating *(SW)*, (vi) dry cough *(DC)*, (vii) loss of appetite *(LAP)*, (viii) rashes *(R)*, (ix) stomachache *(ST)*, (x) constipation or diarrhea *(CD)*, and (xi) delirium *(DL)* [19]. Ten SC then assigned "weights" between "0" and "1" (i.e., float values) on sign symptoms and in turn graded based on their vast CKB as "mild", "moderate", and "severe". For each case, "weights" were assigned based on an arbitrary 3-point scale (*mild* ≤ 0.33, *moderate* $> 0.34 < 0.66$, and *severe* > 0.66), based on their CKB. Corresponding severity grades (G) were labeled as "0", "1", and "2", as "mild", "moderate", and "severe", respectively, by the SC. One sample CRB can be seen as follows (refer to Table 1),

The table refers to one sample case having "severe" fever, headache, weakness, muscle pain, sweating, and dry cough; "moderate" loss of appetite, rashes, and stomachache; and "mild" constipation or diarrhea and delirium have been graded as "severe" typhoid fever. It is also a sample of one of the CRBs, constituted by the SCs.

2.2. Statistical data mining

It is important to test the internal consistency or fidelity of the data before MLC modeling. Thus, the CDB *quality* is checked by computing Cronbach's alpha (alpha) with the below Equation (2) [20].

$$alpha = \frac{(N - \bar{c})}{\bar{v} + (r - 1).\bar{c}} \tag{2}$$

In this equation, "N" refers to the number of scaled data, $\bar{c}$ is the mean of all covariances between the data points, and $\bar{v}$ is the average variance. The Cronbach alpha measure is one of the most important preprocessing steps, as low-quality data may affect the performance of MLC models. The consistency score "alpha" is expressed as a number between 0 and 1, where α 0.8 is considered ideal while $\alpha \leq 0.5$ is deemed "unacceptable" [21].

How the data are distributed is also computed with *Shapiro–Wilk test* [22]. For this dataset, the probability or p-value is <0.05 (CI 95%) signifying that the data are "not" normally distributed, meaning only 5% or less of the time the measures fall beyond 2-standard deviation from the mean value [23]. It is important to note that the real-world biological data are often not normally distributed and hence are complex, skewed, and unpredictable [24].

2.3. VJC development

Ten VJCs developed with ten MLCs or DDSSes are developed in *Python 3.8* using "*scikit learn*" packages preinstalled in *Windows 10 pro*.

- *Linear regression* (LR), which predicts the grade by finding the best-fitting line [10, 25]→ VJC1
- *Logistic regression*, which predicts the probability of the labeled or categorical output variable (i.e., the grade of the illness) [9, 26] (LOR) → VJC2
- *Support vector machine*, which separates different classes (here, grades) based on the hyperplane, creates [27, 28] (SVM) → VJC3
- *Decision tree* (DT), which is a flow diagram of possible decision outcomes in each branch of the tree, and the farthest branch gives the final prediction of a class label (i.e., grade of the illness in this study) [29] → VJC4
- *K-Nearest neighborhood* (KNN), which measures the similarity among "k" data points, and final class prediction is made by "majority voting" to its neighbors [30] → VJC5
- *Linear discriminant analysis* (LDA) works by maximizing the distance between any two classes and minimizing the variation within each class [31] → VJC6
- *Gauss naïve Baye's* (GNB), which works on probabilistic classification [32] → VJC7
- *Support vector classifier* (SVC) → VJC8
- *Random forest* (RF) [33], which predicts based on the "majority voting" of the classifications done by each of the multiple DTs → VJC9, and
- *Multilayer perceptron* (MLP), which works by generating outputs (i.e., predicts classes) through input and hidden layer by multiplying the weight vectors and using nodal transfer functions [34] → VJC10

Working principles of these algorithms are well known and can be found in the original references, cited alongside.

2.4. Data scaling

It is important to note that before developing these DDSSes or VJCs, the raw CDB is preprocessed with *a standard* package of data scaling. The reasons are that

a) MLCs (e.g., LR, LOR, MLP-back propagation) that work with the principle of gradient (θ) descent search require scaling of data (x) as the value of "x" plays a crucial role in the output as can be seen in Equation (3).

$$\theta_j = \theta_j - \alpha \frac{1}{N} \sum_{i=0}^{N} (h_\theta(x^i) - y^i) x_j^{(i)} \tag{3}$$

Table 1. A sample CRB after weight assigned to the signs and symptoms

FV	HD	WK	MS	SW	DC	LAP	R	ST	CD	DL	G
0.9235	0.9951	0.7885	0.7903	0.989	0.0129	0.5451	0.4906	0.5829	0.2238	0.2473	2

In this equation, "α" is the learning rate, and $\frac{1}{N}\sum_{i=0}^{N}(h_\theta(x^i) - y^i)x_j^{(i)}$ is the mean squared error (MSE), and

b) MLCs such as KNN and SVM, which work with the principle of distance-based similarity measures, are also affected by feature values (x_i).
c) Although the remaining MLCs are not affected by the feature scales, still as a process of good practice to develop MLC-based DDSSes, the CDB has been scaled by the *standard scalar* package of Python.

2.5. Model training and handling sampling bias

The train: test dataset is divided into 7:3 across all MLCs with *10-fold cross-validations*. While each fold is used for testing, the remaining ninefold datasets are used for training the model iteratively to reduce sampling bias [35]. It is important to note that instead of 10-fold, 3, 5, and even 20-fold approaches can be used as per the complexity of the model.

2.6. Performance

The measuring parameters, computed with "*scikit learn metrics*" (Equations (4)–(7)) of each DDSS/VJC, are as follows:

$$\text{Accuracy(A)}: \frac{(TP + TN)}{(TP + TN + FP + FN)} \tag{4}$$

where TP, FP, TN, and FN are true and false positive and true and false negative, respectively.

$$\text{Precision(P)}: \frac{TP}{(TP + FP)} \tag{5}$$

$$\text{Recall(R)}: \frac{TP}{(TP + FN)} \tag{6}$$

$$F\text{-score(F)}, \text{which is the weighted average of P and R}: \frac{2(P * R)}{(R + P)} \tag{7}$$

P, *R*, and *F* are measured using *Scikit learn metrics*' *Macro*, *Micro*, and *Weighted* averages [36]. A comparative analysis of the performances of the DDSS/VJC can be seen in Table 2 in the following section. It is important to mention here that the visualization of typhoid grading (mild, moderate, and severe) by the best-performing DDSS/VJC can be seen in Figure 3.

Table 2. Compared performances of VJCs and their respective DDSS algorithms

VJC	DDSS	A	P	R	F	Rank
1	LR	0.45	0.5555	0.4634	0.5052	**9**
2	LOR	0.8	0.7993	0.7978	0.7985	**7**
3	SVM	0.38	0.5406	0.3958	0.4570	**10**
4	DT	0.86	0.8666	0.8666	0.8666	**2**
5	KNN	0.8	0.8063	0.8045	0.8053	**6**
6	LDA	0.8166	0.826	0.812	0.8189	**4**
7	GNB	0.8156	0.816	0.802	0.8109	**5**
8	SVC	0.8333	0.8346	0.8312	0.8328	**3**
9	RF	0.8833	0.8835	0.8833	0.8831	**1**
10	MLP	0.75	0.7477	0.75	0.7488	**8**

3. Results and Discussion

The results obtained from the experiments are shown and discussed in this section.

An alpha value of the CDB is found to be 0.63, which indicates moderate-quality data [21]. The data is of "moderate"-quality clinical data and can be considered good data for experiments.

In this table, the performance of each DDSS (i.e., VJC) has been showcased. It can be noticed that VJC9 (DDSS type: RF) outperforms others, followed by VJC4 (DDSS type: DT), and VJC8 (DDSS type: SVC). Among the rest, LDA (VJC6), GNB (VJC7), KNN (VJC5), LOR (VJC2), and MLP (VJC10) have performed moderately well, whereas MLR (VJC1) and SVM (VJC3) have not performed up to the mark. It does not mean that MLR and SVM MLCs are poor-quality DDSSes. In this set of data, it has not performed well but may perform well in another dataset. The overall performance of the VJCs is appreciable, as clinicians' average perfection in diagnosis is 71% [37], which is even less than the DDSSes (about 74%). However, it is important to note that such numerical comparison does not prove the superiority of DDSSes over the human accuracy of medical diagnosis at all time. For both the human and DDSSes, the complexity of the case and the data, respectively, is the key influencer, which again varies from one set to another.

As known to us, RF is a collection of multiple DTs, and its final output is determined by the output of each of the DTs based on the majority voting [38]. In this work, the author proposes that a clinician may also use a similar approach, i.e., creating DTs (tree-based search) on a random subset of "weighted" sign symptoms (constituting the CDB) to arrive at multiple diagnoses initially, known as "differential diagnosis or DD", where each diagnosis represents one class of disease with its corresponding grade, obtained from each of the DTs based on CRB. DD is the preliminary diagnosis. Here, multiple possibilities are considered. The final diagnosis is known as "provisional diagnosis" (PD), which is based on patients' (i) response to the treatment and (ii) pertinent corroborative findings in the investigation results that decide on the "majority voting" of the most likely "type" of the disease and its "grade" within the list of DD. This is the rule-of-the-thumb method of medical diagnosis that is quite similar to the working principle of the DT-RF DDSS that makes decisions based on the classes obtained from the respective DTs. Figure 2 showcases the similarity between DT-RF-based DDSS and a clinician's method of diagnosis and grading of any disease. This method is mostly used in cases that mandate a faster PD, such as in the diagnosis of IDs and their grades at the time of diagnosis to initiate early treatment and prevent its further progression in the community. As DT- and RF-based DDSS have faster convergence, these probably have outperformed other DDSS algorithms in this study. In Figure 2, INSTANCES are the weighted sign symptoms, i.e., the CDB; a DT-based CLASS is nothing but one of the diagnoses in the list of DDs, which is done with CRB; RF-based "majority voting" leading to FINAL CLASS is nothing but the PD or final diagnosis of the grade of illness. In this way, DT-RF-based DDSS (an ensemble model) may model how a novice clinician learns to diagnose the grade of an illness.

3.1. Visualization

Predicted grades of the test cases by the RF-based DDSS, i.e., VJC9 are plotted below. In this figure, purple markers are "mild" grades, green markers are "moderate" grades and yellow markers denote "severe" grades of typhoid fever in 60 test cases.

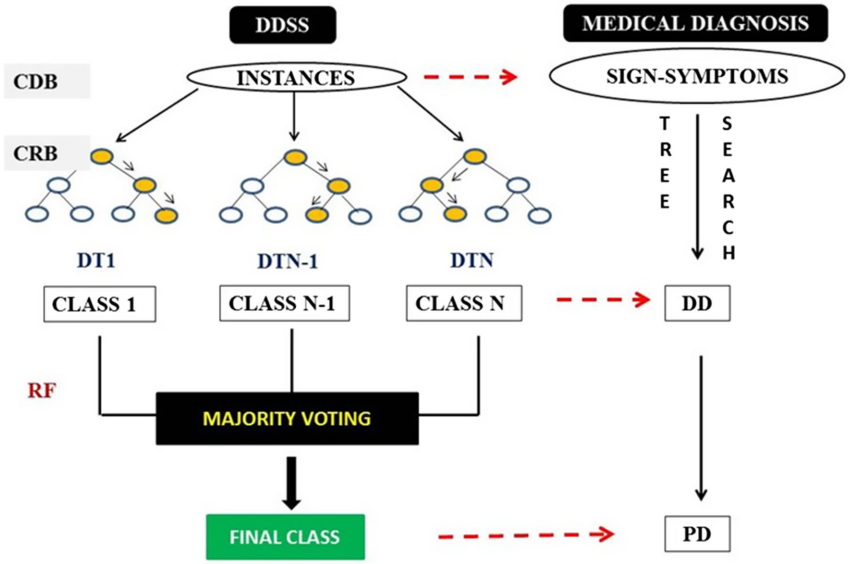

Figure 2. Similarity between DT- and RF-based DDSS and medical diagnosis

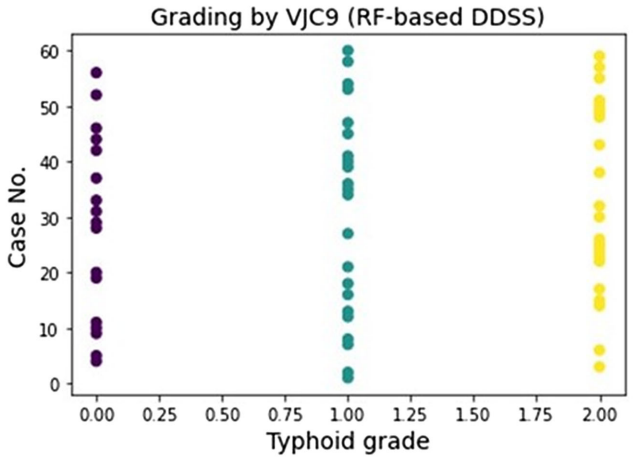

Figure 3. Grading by VJC9 (RF-based DDSS)

Screenshots of the results (accuracy, precision, recall, and *F*-score) from the *IPython console* of *Spyder 5.0 editor* have been shown in the figure for the test cases.

4. Challenges and Hardships with MLCS as CDSS

Clinical decision- making is essentially a "human-centric" process where "human" patients present with signs and symptoms (i.e., the patterns) of an illness that are analyzed by the "human" doctors based on the knowledge they have acquired dynamically over the years of clinical practice (i.e., their experience enriched with CKB) both during the training phase (supervised learning and reinforcement learning) and own-run independent practice (unsupervised learning). Computer-aided decision making is usually needed to analyze complex and difficult cases by experienced doctors, while novice doctors may use those to obtain some prior knowledge in plausible diagnoses, grades, recommendations, and prognoses. The interesting point is that computers can remember and recall thousands of different patterns in a short time, whereas the human brain often forgets and falters but can assume, imagine, and execute the most appropriate clinical decisions that computers still cannot. Thus, the involvement of both humans and machines (a hybrid model) could be beneficial to patients (moving ahead from patient-centric to patient-centered) in this digital era of healthcare. However, there are several challenges to accomplishing it.

a) *Explainability and causability* are two important aspects of model's *interpretability* bringing logical relationships between the factors and the response more toward human acceptable format than statistical and numeric formats, which sometimes do not make enough sense to the clinicians. The "black box" algorithms are confusing to them.
b) *Generalizability* is another issue as MLCs are trained and tested on a dataset and do not always show a similar level of precision and accuracy in other real-life scenarios. For example, the CDSS's performance may vary between two or more hospital setups in the same domain and disease spectrum.
c) *Multimodal data handling* is another important hardship as many come from IoT-enabled devices, e.g., ICU or ITU patient data. Multiple X-rays, CT, MRI, USG, and other different types of images come from PACS. Laboratory data, data from pharmacies, and doctors' and nurses' notes on procedures and referrals make it more complex, especially if those are hand-written. All this information is important to corroborate the signs and symptoms of the patients to stratify the clinical risk in real time and plan care management and treatment accordingly.
d) *Noise handling* is another important aspect of healthcare data, which is usually unstructured, scattered, and noisy.
e) *Data imbalance* (abundance of one particular class type in the dataset) is one critical parameter hindering model's performance.
f) *Ethical measures* are key to healthcare practices. Data privacy is the most important aspect of it. The popular choice of using OpenAI such as LLMs is still questionable because the data-analyzing platforms requires critical thinking that is absent within it. Moreover, there are incidences of the data leak, phishing, and generating wrong information (hallucinated information) [39], violations of data security norms and compliances while gathering sensitive personal data of the patients [40], and nonrepeatability with the same user input

confuses the users further. There are similar threats to the IoT and image data also despite HL7 and DICOM formatting for archiving and transmission. Often, security is breached by malware and spyware [41]. An attempt has to be made to keep the data anonymous at all times, masking its sources, encrypting, and tightening the security. Therefore, LLMs and OpenAI platforms have questionable suitability to handle sensitive healthcare data.

The plausible *solutions* to the above challenges could be as follows:

i) *Data wrangling* methods such as preprocessing and processing are the most useful techniques before model building. Column-wise normalization, identifying null values, and data types are very important too to successfully develop an MLC model.
ii) *Training* one model at a time with several datasets coming from different sources, e.g., the typhoid fever dataset from ten different hospitals to reduce the inter-rater variance among the attributes in the datasets and the skewness in the output.
iii) *Stratified k-fold cross-validation* helps curb the sampling bias which otherwise is carried forward to influence the outcome of the model.
iv) *Validation* and *cross-validation* of the outputs by a dedicated team of experienced doctors (double-blind controlled trial) whenever possible and involving them at all stages of the model development for syncing CKB with the algorithm's speed and memory.
v) *Type-I and II error handling* is an important metric of the performance of any CDSS tools. It should be measured to the occurrences of false-positive (Type-I error) and false-negative cases (Type-II error). The intention must be around reducing these scores as false-negative results could entail significant health hazards as the clinicians could overlook the imminent risk, while false positivity may push clinicians for more high-end investigations and allocation of sophisticated resources, which may raise the cost to the patients.
vi) *Exploratory data analysis* meticulously mines the hyperparameters and how they influence each other as well as the model's output. It presents how much weight each of these hyperparameters is bringing in addressing the model's outcome.
vii) Making the CDSS more *human-centric*, *clinically rich*, and must be *user-friendly* for clinical adoption.
viii) *Retesting the model's performance* in different clinical environments and datasets of the same clinical domain, and
ix) In the case of rule-based systems, most of the cases generate thousands of "if-then" rules, which need to be pruned to identify the "useful" bunch mimicking how the clinicians arrive at a diagnosis by excluding apparently "useless" or nonpertinent rules [14]. *Rule pruning* reduces computational complexities, increases the model accuracy, and increases the speed of diagnostic decision making.

5. Contributions

The contribution of this paper is three pronged:

A. It effectively develops an ML-based CDSSes for diagnosing a complex systemic illness such as typhoid fever. All models are developed with the weighted signs and symptoms of the patients, given by a group of experienced clinicians meaning these are trained with the rich CKB mimicking the way novice doctors are trained by experienced senior doctors. It essentially gives the essence and core idea behind using CDSS for disease diagnosis and grading.
B. It technically maps how human doctors diagnose using the "yes" and "no" heuristics logic in arriving at a diagnostic decision, which is an advantage of using the DT-RF-CDSS models in the medical diagnoses, and
C. By incorporating the essence of computations, such a trained CDSS, can assist novice medical doctors learning diagnostic methods for complex illnesses, while experienced doctors may have a second opinion with the tool. The use of computation is to recapitulate all the closely related disease patterns within the signs and symptoms, which the brain struggles to decode with age.

6. Limitation

A. The data are synthetic and not the real world, though it is created by experienced general physicians who treat typhoid fever cases. The author thus believes that the real-world data will be closely around the synthetic dataset having an appreciable internal consistency (alpha score of 0.63). Also, it is not normally distributed, corroborating with the fact that biological data are often not normally distributed and hence are unpredictable. Therefore, apparently, it is a limitation from the research perspective but when critically thought, it is not.
B. The sample size is small and might have hindered better outcomes of the ML algorithms. However, it is not always correct. The quality or relevance of the data is more important for ML modeling than its size, e.g., if the sampling happens from wrong distribution having a wrong mean, more data help converging faster toward that wrong mean affecting the model's outcome extensively. In this case, the data signify the CKB of the SCs having decades of experience in treating IDs. Therefore, it is assumed that the quality of the data, especially the feature set is very high although being smaller in size, and it serves the purpose of the key objective of the paper which is how the learning of ML algorithms for clinical decision-making mimic the way novice doctors learn diagnose and make clinical decisions. High-quality feature set is the key to building a successful MLC model [42].

7. Conclusion

The paper attempts to mathematically model how a novice clinician (a VJC) learns to diagnose a complex illness that requires a faster decision, such as typhoid fever using the CKB of the SC. To accomplish the goal, ten VJCs are created using ten MLCs for grading typhoid fever. The VJCs are trained by the SC to make a faster tree-search-based approach to preliminarily decide on the disease grade based on the "weighted" sign symptoms of the patients. Finally, for a new case to be diagnosed, majority votes decide the grade of the illness, learned from SC's CRB. That is why, VJCs, developed with DT-RF-MLC-based CDSS, have performed best in diagnosing the grade of the illness with >87% average accuracy and precision, which is higher than human clinician's reported diagnostic accuracy, i.e., 71% [37]. The paper also provides insight into selecting the appropriate MLC algorithms for the diagnostic decision-making process.

At the experimental level, it looks simpler. However, interoperability is a challenge for the deployment and adoption of MLCs by human doctors as assistive tools, and therefore, the models need extensive testing across other datasets of different sources. Finally, the performance can be tested majorly based on the frequencies of occurrence of Type-I (number of false-positive cases) and II (number of false-negative cases) errors in domain-specific real-world diagnostic settings.

Ethical Statement

This study does not contain any studies with human or animal subjects performed by the author.

Conflicts of Interest

The author declares that he has no conflicts of interest in this work.

Data Availability Statement

The data that support this work are available upon reasonable request to the corresponding author.

Author Contribution Statement

Subhagata Chattopadhyay: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing - original draft, Writing - review & editing, Visualization.

References

[1] World Health Organization. (2024). *Immunization, vaccines and biologicals.* Retrieved from: https://www.who.int/teams/immunization-vaccines-and-biologicals/diseases/typhoid#:~:text=According%20to%20the%20most%20recent,(or%20uncommonly%20Paratyphi%20C)

[2] Bhan, M. K., Bahl, R., & Bhatnagar, S. (2005). Typhoid and paratyphoid fever. *The Lancet*, *366*(9487), 749–762. https://doi.org/10.1016/S0140-6736(05)67181-4

[3] Dougan, G., & Baker, S. (2014). *Salmonella enterica* serovar Typhi and the pathogenesis of typhoid fever. *Annual Review of Microbiology*, *68*, 317–336. https://doi.org/10.1146/annurev-micro-091313-103739

[4] Ahmed, A., & Ahmed, B. (2010). Jaundice in typhoid patients: Differentiation from other common causes of fever and jaundice in the tropics. *Annals of African Medicine*, *9*(3), 135–140. https://doi.org/10.4103/1596-3519.68361

[5] Parry, C. M., Hoa, N. T. T., Diep, T. S., Wain, J., Chinh, N. T., Vinh, H., . . . , & Farrar, J. J. (1999). Value of a single-tube Widal test in diagnosis of typhoid fever in Vietnam. *Journal of Clinical Microbiology*, *37*(9), 2882–2886. https://doi.org/10.1128/jcm.37.9.2882-2886.1999

[6] Voysey, M., Pant, D., Shakya, M., Liu, X., Colin-Jones, R., Theiss-Nyland, K., . . . , & Pitzer, V. E. (2020). Under-detection of blood culture-positive enteric fever cases: The impact of missing data and methods for adjusting incidence estimates. *PLOS Neglected Tropical Diseases*, *14*(1), e0007805. https://doi.org/10.1371/journal.pntd.0007805

[7] Mogasale, V., Ramani, E., Mogasale, V. V., & Park, J. (2016). What proportion of *Salmonella* Typhi cases are detected by blood culture? A systematic literature review. *Annals of Clinical Microbiology and Antimicrobials*, *15*, 32. https://doi.org/10.1186/s12941-016-0147-z

[8] Mawazo, A., Bwire, G. M., & Matee, M. I. N. (2019). Performance of Widal test and stool culture in the diagnosis of typhoid fever among suspected patients in Dar es Salaam, Tanzania. *BMC Research Notes*, *12*, 316. https://doi.org/10.1186/s13104-019-4340-y

[9] Chattopadhyay, S., Chattopadhyay, A. K., & Aifantis, E. C. (2021). Predicting case fatality of dengue epidemic: Statistical machine learning towards a virtual doctor. *Journal of Nanotechnology in Diagnosis and Treatment*, *7*, 10–24. https://doi.org/10.12974/2311-8792.2021.07.2

[10] Chattopadhyay, A. K., & Chattopadhyay, S. (2022). VIRDOCD: A VIRtual DOCtor to predict dengue fatality. *Expert Systems*, *39*(1), e12796. https://doi.org/10.1111/exsy.12796

[11] Chattopadhyay, S. (2013). Mathematical modelling of doctors' perceptions in the diagnosis of depression: A novel approach. *International Journal of Biomedical Engineering and Technology*, *11*(1), 1–17. https://doi.org/10.1504/IJBET.2013.053702

[12] Procaccianti, M., Motta, A., Giordani, S., Riscassi, S., Guidi, B., Ruffini, M., . . . , & Dodi, I. (2020). First case of typhoid fever due to extensively drug-resistant *Salmonella enterica* serovar Typhi in Italy. *Pathogens*, *9*(2), 151. https://doi.org/10.3390/pathogens9020151

[13] Centers for Disease Control and Prevention. (2020). *Typhoid fever and paratyphoid fever: Vaccination.* Retrieved from: https://www.cdc.gov/typhoid-fever/typhoid-vaccination.html#:~:text=Typhoid%20vaccines%20are%20not%20100,a%20booster%20every%205%20years

[14] Ashish, K., Dasari, A., Chattopadhyay, S., & Hui, N. B. (2018). Genetic-neuro-fuzzy system for grading depression. *Applied Computing and Informatics*, *14*(1), 98–105. https://doi.org/10.1016/j.aci.2017.05.005

[15] Bentaiba-Lagrid, M. B., Bouzar-Benlabiod, L., Rubin, S. H., Bouabana-Tebibel, T., & Hanini, M. R. (2020). A case-based reasoning system for supervised classification problems in the medical field. *Expert Systems with Applications*, *150*, 113335. https://doi.org/10.1016/j.eswa.2020.113335

[16] Ayo, F. E., Awotunde, J. B., Ogundokun, R. O., Folorunso, S. O., & Adekunle, A. O. (2020). A decision support system for multi-target disease diagnosis: A bioinformatics approach. *Heliyon*, *6*(3), e03657. https://doi.org/10.1016/j.heliyon.2020.e03657

[17] Oguntimilehin, A., Olatunji, K. A., & Abiola, O. B. (2020). A computer based intelligent system for managing typhoid fever. *International Journal of Advanced Trends in Computer Science and Engineering*, *9*(1), 528–532. https://doi.org/10.30534/ijatcse/2020/72912020

[18] Oguntimilehin, A., Adetunmbi, A. O., & Abiola, O. B. (2013). A machine learning approach to clinical diagnosis of typhoid fever. *A Machine Learning Approach to Clinical Diagnosis of Typhoid Fever*, *2*(4), 1–6. http://eprints.abuad.edu.ng/73/

[19] Centers for Disease Control and Prevention. (2024). *Symptoms of typhoid fever and paratyphoid fever*. Retrieved from: https://www.cdc.gov/typhoid-fever/signs-symptoms/?CDC_AAref_Val=https://www.cdc.gov/typhoid-fever/symptoms.html

[20] Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, *16*(3), 297–334. https://doi.org/10.1007/BF02310555

[21] Goforth, C. (2015). *Using and interpreting Cronbach's Alpha.* Retrieved from: https://library.virginia.edu/data/articles/using-and-interpreting-cronbachs-alpha

[22] Shapiro, S. S., & Wilk, M. B. (1965). An analysis of variance test for normality (complete samples). *Biometrika*, *52*(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

[23] Cardinal, L. J. (2015). Central tendency and variability in biological systems. *Journal of Community Hospital Internal Medicine Perspectives*, *5*(3), 27930. https://doi.org/10.3402/jchimp.v5.27930

[24] Mar, J. C. (2019). The rise of the distributions: Why non-normality is important for understanding the transcriptome and beyond.

Biophysical Reviews, *11*(1), 89–94. https://doi.org/10.1007/s12551-018-0494-4

[25] Kenney, J. F., & Keeping, E. S. (1962). Linear regression and correlation. In J. F. Kenney (Ed.), *Mathematics of statistics* (pp. 252–285). D. Van Nostrand Company.

[26] Walker, S. H., & Duncan, D. B. (1967). Estimation of the probability of an event as a function of several independent variables. *Biometrika*, *54*(1–2), 167–179. https://doi.org/10.1093/biomet/54.1-2.167

[27] Boser, B. E., Guyon, I. M., & Vapnik, V. N. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, 144–152. https://doi.org/10.1145/130385.130401

[28] Cortes, C., & Vapnik, V. (1995). Support-vector networks. *Machine Learning*, *20*, 273–297. https://doi.org/10.1007/BF00994018

[29] Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning*, *1*, 81–106. https://doi.org/10.1007/BF00116251

[30] Fix, E., & Hodges, J. L. (1951). *Discriminatory analysis: Nonparametric discrimination: Consistency properties*. USA: USAF School of Aviation Medicine.

[31] Fisher, R. A. (1936). The use of multiple measurements in taxonomic problems. *Annals of Eugenics*, *7*(2), 179–188. https://doi.org/10.1111/j.1469-1809.1936.tb02137.x

[32] Stigler, S. M. (1983). Who discovered Bayes's theorem? *The American Statistician*, *37*(4a), 290–296. https://doi.org/10.1080/00031305.1983.10483122

[33] Breiman, L. (2001). Random forests. *Machine Learning*, *45*, 5–32. https://doi.org/10.1023/A:1010933404324

[34] Rosenblatt, F. (1962). *Principles of neurodynamics: Perceptrons and the theory of brain mechanisms*. USA: Spartan Books.

[35] Prusty, S., Patnaik, S., & Dash, S. K. (2022). SKCV: Stratified K-fold cross-validation on ML classifiers for predicting cervical cancer. *Frontiers in Nanotechnology*, *4*, 972421. https://doi.org/10.3389/fnano.2022.972421

[36] Scikit Learn. (n. d.). *fi_score*. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.metrics.f1_score.html

[37] Richens, J. G., Lee, C. M., & Johri, S. (2020). Improving the accuracy of medical diagnosis with causal machine learning. *Nature Communications*, *11*(1), 3923. https://doi.org/10.1038/s41467-020-17419-7

[38] Salini, A., & Jeyapriya, U. (2018). A majority vote based ensemble classifier for predicting students academic performance. *International Journal of Pure and Applied Mathematics*, *118*(24), 1–11. https://acadpubl.eu/hub/2018-118-24/1/6.pdf

[39] Pool, J., Akhlaghpour, S., Fatehi, F., & Burton-Jones, A. (2024). A systematic analysis of failures in protecting personal health data: A scoping review. *International Journal of Information Management*, *74*, 102719. https://doi.org/10.1016/j.ijinfomgt.2023.102719

[40] Nicastro, D. (2023). *OpenAI faces lawsuit over stealing millions of personal consumer data*. Retrieved from: https://www.cmswire.com/digital-experience/openai-faces-lawsuit-over-stealing-millions-of-personal-consumer-data/

[41] Eichelberg, M., Kleber, K., & Kämmerer, M. (2020). Cybersecurity challenges for PACS and medical imaging. *Academic Radiology*, *27*(8), 1126–1139. https://doi.org/10.1016/j.acra.2020.03.026

[42] Morán-Fernández, L., Bólon-Canedo, V., & Alonso-Betanzos, A. (2022). How important is data quality? Best classifiers vs best features. *Neurocomputing*, *470*, 365–375. https://doi.org/10.1016/j.neucom.2021.05.107

How to Cite: Chattopadhyay, S. (2024). Decoding Medical Diagnosis with Machine Learning Classifiers. *Medinformatics*. https://doi.org/10.47852/bonviewMEDIN42022583

Abbreviations

BDSS	bioinformatics-based decision system
CD	constipation or diarrhea
CDB	clinical database
CDSS	Clinical Decision Support System
CRB	clinical rule base
CKB	clinical knowledge base
DC	dry cough
DD	differential diagnosis
DDSS	Diagnostic Decision Support System
DL	delirium
DT	decision tree
FV	fever
GNB	Gauss naïve Baye's
HD	headache
ID	infectious disease
KNN	K-nearest neighborhood
LAP	loss of appetite
LDA	linear discriminant analysis
LLM	large language models
LR	linear regressions
LOR	logistic regressions
MDR	multidrug resistant
MLC	machine learning classifiers
MS	muscle pain
MLP	multilayer perceptron
PACS	Picture Archiving and Communication System
PD	provisional diagnosis
R	rashes
RF	random forest
SC	senior clinicians
ST	stomachache
SVC	support vector classifier
SVM	support vector machine
SW	sweating
VJC	virtual junior clinician
WK	weakness.