Intelligence-Based Medicine 5 (2021) 100029

Contents lists available at ScienceDirect

Intelligence-Based Medicine

journal homepage: www.journals.elsevier.com/intelligence-based-medicine

Predicting shunt infection in children with hydrocephalus

M. Sabeti [a], R. Boostani [b], E. Moradi [c,*], Z. Habibi [d], F. Nejat [d]

[a] Department of Computer Engineering, North Tehran Branch, Islamic Azad University, Tehran, Iran
[b] CSE & IT Dept., Faculty of Electrical and Computer Engineering, Shiraz University, Shiraz, Iran
[c] Department of Neurosurgery, Shahid Beheshti University of Medical Sciences, Tehran, Iran
[d] Department of Neurosurgery, Tehran University of Medical Sciences, Tehran, Iran

ARTICLE INFO

Keywords:
Hydrocephalus
Shunt infection
Machine learning
Prediction

ABSTRACT

Hydrocephalus is defined as the increase in Cerebro Spinal Fluid (CSF) volume, which is usually accompanied by high Intracranial Pressure (ICP). The most common treatment for hydrocephalus is ventriculoperitoneal shunt insertion. Shunt is a tube which drains CSF from the ventricular system to peritoneal cavity. Then, the CSF is absorbed from peritoneum. Infection is considered as one of the most complications of shunt systems, which can cause improper prognosis in patients, especially in children's neuro development. Hence, identifying shunt infection predictive factors could improve the practice in preventing this event. This study used a dataset containing the features of 68 patients with a history of shunt infection and 80 patients without any history of shunt infection (control group) in Children's Medical Center hospital of Tehran (Iran). The state-of-art techniques were applied to select the most informative predicting factors (features). The probability (accuracy) of shunt infection was determined with different intelligent and statistical classifiers. The results indicated that history of prematurity and intraventricular hemorrhage, age of the first shunt procedure, number of shunt revisions, brain tumor induced hydrocephalus, birth weight, and coinfection are the best descriptive features. In addition, the best classification results by different techniques varied in the accuracy range of 68%–81% in the dataset.

1. Introduction

Cerebro Spinal Fluid (CSF) is continuously made by Central Nervous System (CNS), which carries nutrients, washes away impurities, and acts as a cushion for CNS. After production, it flows through ventricles and sub-arachnoid space and is finally absorbed by brain venous blood stream. A precise balance between the production and absorption of CSF is necessary to maintain normal Intra-Cranial Pressure (ICP). There is an excessive amount of CSF in hydrocephalus intradural space, which could be related to an obstruction in CSF flow or a defect in absorption to brain blood flow [1].

In general, it occurs in 1–2 per 100 live births in children and adults. Several different situations such as tumors, infections, trauma, developmental abnormalities, and other factors can cause hydrocephalus. Hydrocephalus, as a neurosurgical emergency, increases head circumstance, decreases vision, less of consciousness and some other neurodevelopmental deficits. Ventricular shunt placement is considered as the most common treatment for symptomatic hydrocephalus. Shunt is an internal tube which drains CSF from ventricles to other places in the body in order to absorb the extra fluid and preserve normal ICP [2,3].

Shunt has several complications despite its numerous benefits, the most common complication of which is its infection, which has many morbidities, especially in the neurodevelopmental growth of the affected children and enormous economic costs for healthcare systems. The rate of shunt infection varies from 10% to 22%, and approximately 90% of the infections usually occur one month after surgery. There are some potential predisposing factors for shunt infection such as the patient's age, etiology of hydrocephalus, hospitalization period, number of shunt revisions, surgeon's experience, operation duration, surgical technique, manipulation of the indwelling device during surgery, and health insurance [4].

Neurosurgeons use some customized protocols to reduce the shunt infection rate. However, shunt infection is still the most important complication of hydrocephalus treatment. Hence, the identification of its predictive factors can improve current practices in preventing this catastrophic event. Some medical informatics such as Logistic Regression (LR) and Artificial Neural Networks (ANNs) have recently been applied to develop models for the prediction task. Habibi et al. [5] studied 68 patients with shunt infection and 80 controls which fulfilled a set of meticulous inclusion/exclusion criteria. They performed univariate

* Corresponding author.
E-mail address: moradieh@sbmu.ac.ir (E. Moradi).

https://doi.org/10.1016/j.ibmed.2021.100029
Received 13 September 2020; Received in revised form 19 January 2021; Accepted 25 February 2021

analysis for a long list of risk factors, and those with p-value< 0.2 were used to create ANN and LR models. Five variables including birth weight, age of the first shunting procedure, shunt revision, prematurity history, and myelomeningocele were significantly associated with shunt infection via univariate analysis, and two other variables (intraventricular hemorrhage & coincided infections) had a p-values of less than 0.2. The results showed that ANN and LR models could predict shunt infection with an accuracy of 83.1% and 55.7%, respectively. Based on the findings, ANN could predict shunt infection with a fairly high level of accuracy in children with shunted hydrocephalus.

In general, infection management is a highly complex issue. For example, life-threatening conditions such as sepsis require immediate diagnostic and treatment while the causative pathogen is often unknown. Luz et al. [6] investigated different applications of Machine Learning (ML) for clinical decision support in infectious diseases to support diagnosis, predict severity, and choose appropriate antimicrobial treatments. The results indicated that early identification of septic patients through ML-derived prediction models could improve and facilitate patient care in situations where time is life.

Data collection plays an important role in the diagnosis and treatment of patients. Physicians should cope with a large number of subject-dependent factors and review the patient's whole history. Data are mainly collected to find out the relevant patterns in the investigated disease. During the past two decades, various data mining schemes have become highly important in diagnosing and treating different diseases [7, 8]. Raghavendra et al. [9] presented a review of research on automated diagnosis of five neurological disorders using ML techniques on the elicited features from physiological signals and images. They investigated some diseases such as epilepsy, Parkinson, Alzheimer, Multiple Sclerosis (MS), and ischemic brain stroke. In addition, they reviewed recent research articles by focusing on their feature extraction methods, dimensionality reduction techniques, feature selection schemes, and classification techniques. They concluded that the integration of ML techniques in an automated fashion could assist neurologists, neurosurgeons, radiologists, and other medical providers to make better clinical decisions.

Azimi et al. [10,11] developed an ANN model to predict Endoscopic Third Ventriculostomy (ETV) success at 6 months and compared the findings to those obtained by traditional predictive measures in childhood hydrocephalus. They examined the data collected from 168 patients (80 males & 88 females; mean age: 1.4 ± 2.6 years) and applied ANN, ETV Success Score, CURE children's hospital of Uganda ETV success score, and LR models for the prediction task. Further, they considered various features such as hydrocephalus causes, age, sex, Choroid Plexus Cauterization (CPC), previous shunt surgery, hydrocephalus type, and body weight. The results showed that etiology, age, CPC status, hydrocephalus type, and previous shunt placement are the most important features. Furthermore, ANN models could produce better results with an accuracy rate of 95.1% and an area under the curve of 0.87 in comparison with the other models.

Neurosurgeons have applied many standardized protocols to lower the occurrence of shunt infection for many years. However, shunt infection is still the most significant and prevalent complication associated with hydrocephalus treatment causing serious problems for the affected children, their family, and the healthcare system. Identifying children with higher risks of shunt infection can significantly improve the management of this situation. Previous studies indicated that many different risk factors are related to either patients or surgical aspects. However, clinical predictors of shunt infection among children are still controversial. The present study aimed to investigate the performance of the state-of-the-art ML techniques to predict shunt infection in hydrocephalus children.

2. Materials

As mentioned earlier, hydrocephalus is defined as the increased amount of CSF in the CNS, which usually happens as a result of obstruction in CSF pathway or decrease in CSF absorption. Fig. 1 shows a hydrocephalus brain in comparison with a normal brain. Shunt infection is known as the identification of a bacterial pathogen in CSF or shunt hardware culture. In children with negative CSF culture and clinical evidence of CNS infection, shunt infection is considered when the CSF analysis parameters are abnormal, as well as during the exposure of shunt device or presence of infected pseudocyst in abdomen. Abnormal CSF parameters include positive smear, low glucose level (<40 mg/dL), and high white blood cell count (>10 cells/mm3) with polymorphonucleosis.

In this study, 148 hydrocephalus patients were selected based on a set of objective inclusion/exclusion criteria among more than 800 ventriculoperitoneal shunt procedures performed by the senior author [5] in Children's Medical Center Hospital of Tehran (Iran) on hydrocephalus patients under the age of 12. A total of 68 patients with shunt infection were consecutively enrolled while 80 patients without shunt infection with the same protocol & inclusion/exclusion criteria were considered as controls. The patients were included only if they had undergone ventriculoperitoneal shunting in an elective setting with a standard protocol and completed a follow-up period of at least 6 months. The method and time of surgery, prophylactic antibiotic, operation theater settings, and the number of staff inside the theater were equal in all cases. Those with ventriculo-atrial shunting, operation in an emergent setting, first procedure in other centers, deviation from the protocol, incomplete or inaccessible medical data, and incomplete or missing follow-up were excluded from the study. For each patient, demographic and medical information including sex, parents' consanguinity, gestational age at birth, delivery type, birth weight, prematurity, head circumference at birth, neonatal icterus, myelomeningocele (MMC) history, meningitis history, intraventricular hemorrhage (IVH) history, head trauma, brain tumor, age of surgery time, surgery duration, type of inserted shunt, other-site active infection within 30 days prior to shunt insertion, CSF leak after shunting, and the number of previous shunt revisions were recorded. Table 1 shows the recorded features in this study.

3. Methods

The present study aimed to implement the model applying ML techniques based on 11 variables. Fig. 2 shows the simple flowchart of this study, and the techniques used are described in the following subsections. As displayed in Fig. 2, we tried to learn from shunt infection observations in the training phase. Then, the proposed approach predicts the risk of shunt infection in new hydrocephalus patients (never-seen before patients). Since feature selection by human mind is just feasible when we face with a few features, an automatic feature selection technique is seriously needed to select important factors in the case of encountering with a large number of features.

3.1. Feature selection

Feature selection was used for reducing dimensionality, and the relevant features were selected while irrelevant and redundant ones were discarded [13]. A reduction in feature dimensionality can improve the prediction performance since it decreases the model complexity and increases its generalization capacity. Different medical applications have been studied by using various types of ML approaches [7,8]. However, to the best of our knowledge, no study has focused on analyzing the effectiveness of feature selection to predict the shunt infection among children with hydrocephalus. Hence, various feature dimensionality reduction methods, as well as the feature weighting technique, were applied in this study. The most commonly used feature selection strategies are Sequential Feature Selection (SFS) [14] L-Plus R-Minus [15], Least Absolute Shrinkage and Selection Operator (LASSO) [16], evolutionary methods (e.g., genetic algorithm) [17], Greedy Overall Relevancy (GOR) [18], and well-known dimensionality reduction methods such as Principal Component Analysis (PCA) and Linear Discriminant Analysis (LDA)

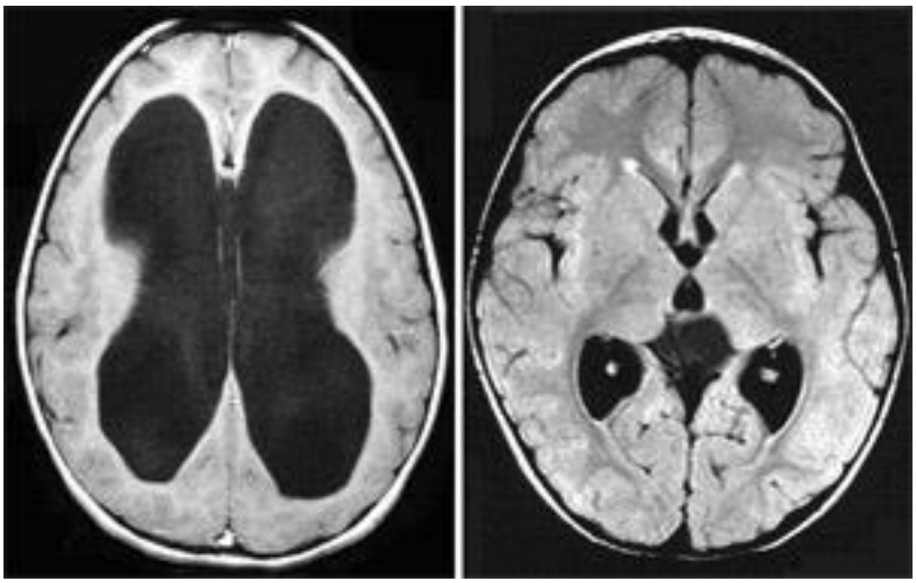

Fig. 1. The hydrocephalic brain (left) & the normal brain (right) [12].

Table 1
The available features from 148 children with hydrocephalus.

Risk factors	With shunt infection	Without shunt infection	Total
Birth weight			
≤ 2000	14	5	19
2000-2500	4	6	10
2500-3000	28	23	51
3000-3500	13	37	50
≥ 3500	9	9	18
Prematurity			
Yes	25	14	39
No	43	66	109
Trauma			
Yes	0	2	2
No	68	78	146
History of meningitis			
Yes	3	3	6
No	65	77	142
Tumor			
Yes	3	5	8
No	65	75	140
History of IVH			
Yes	13	10	23
No	55	70	125
Icter			
Yes	9	4	13
No	59	76	135
coinfection			
Yes	6	3	9
No	62	77	139
History of MMC			
Yes	9	24	33
No	59	56	115
Age at first surgery			
< 2 *weeks*	12	1	13
> 2 *weeks*	56	79	135
Number of shunt revisions			
≤ 4	64	80	144
> 4	4	0	4

[19]. The motivation behind all the aforementioned algorithms is to automatically select an informative subset of features which is most relevant to the prediction task.

3.1.1. Sequential Feature Selection

SFS algorithm is a greedy search algorithm used to project an initial feature space (d-dimensional) to the secondary feature space (k-dimensional) where k < d. In SFS, the classifier accuracy can be selected as the fitness value providing a feedback to the SFS algorithm for selecting the most relevant feature at each epoch. This process continues until the predefined number of features is selected (here, k is set to 7 through the cross validation). Fig. 3 presents the whole procedure of SFS algorithm.

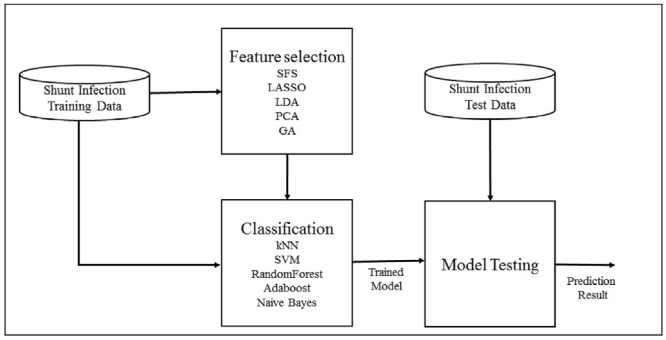

Fig. 2. The simple flowchart of prediction task.

1. Initialize the feature subset with an empty set $Y_k = \varnothing$, $k = 1$.
   2.Select the next best feature as follows

$$x^+ = argmax\, J(Y_k + x)$$

where x^+ is the feature is associated with the best classifier performance if it is added to the feature subset Y_k.

3. Add an additional feature x^+ to the feature subset X_k,

$$Y_{k+1} = Y_k \cup x^+$$

4. Repeat this procedure until k features are selected.

3.1.2. Least Absolute Shrinkage and Selection Operator

The Least Absolute Shrinkage and Selection Operator (LASSO) [20] aims to improve the prediction accuracy and interpretability of the regression models by altering the model fitting process to select a relevant subset of features. LASSO penalizes the absolute norm of the regression coefficients based on the value of a regularization parameter λ.

1. Initialize the feature subset with an empty set $Y_k = \emptyset$, $k = 1$.
2. Select the next best feature as follows
$$x^+ = argmax\, J(Y_k + x)$$
where x^+ is the feature is associated with the best classifier performance if it is added to the feature subset Y_k.
3. Add an additional feature x^+ to the feature subset X_k,
$$Y_{k+1} = Y_k \cup x^+$$
4. Repeat this procedure until k features are selected.

Fig. 3. The overall description of SFS algorithm.

LASSO can be especially useful in selecting variables when there are several possible predictors, most of which show zero or little influence on a target variable.

3.1.3. Principal Component Analysis

PCA [21] is a dimensionality reduction technique which converts initial features by multiplying them into a linear matrix in order to achieve a reduced number of informative features in a shunt infection dataset with minimum loss of information. PCA is performed by selecting eigen-vectors of covariance matrix of the data corresponding to the largest eigen-values, followed by multiplying samples to these selected vectors. Fig. 4 describes the PCA algorithm in details.

1. Standardize the dataset X with subtracting the mean value of samples from X

$$\mu = \frac{1}{N}\sum_{i=1}^{N} x_i \tilde{X} = X - \mu$$

2. Compute the covariance matrix as

$$\Sigma = \frac{1}{N}\tilde{X}^T\tilde{X}$$

3.Calculate the eigenvectors and eigenvalues of Σ
4. Pick $d^{'}$ eigenvectors corresponding to the largest eigen values and put them in the column of $A = [v_1, \ldots, v_{d^{'}}]$
5. $X^{'} = A^T X$

3.1.4. Linear Discriminant Analysis

LDA [22] is closely related to PCA since both of them look for linear combinations of input features into a secondary space with lower dimensions. LDA linearly projects input features into a hyperplane, on which the Fisher criterion is maximized, which results in maximizing the between class scatter matrix while simultaneously minimizing the within class scatter matrix. Fig. 5 shows the overall description of LDA feature selection.

1. Suppose two classes have mean values of μ_1 *and* μ_2 and covariances of Σ_1 and Σ_2
Estimate the between-groups scatter matrix S_B as $S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$ Estimate the within-groups scatter matrix S_W as
$S_W = S_1 + S_2\; S_1 = \sum_{x_i \varepsilon C_1} (x_i - \mu_1)(x_i - \mu_1)^T\; S_2 = \sum_{x_i \varepsilon C_2} (x_i - \mu_2)(x_i - \mu_2)^T$

1. Standardize the dataset X with subtracting the mean value of samples from X
$$\tilde{X} = X - \mu, \quad \mu = \frac{1}{N}\sum_{i=1}^{N} x_i$$
2. Compute the covariance matrix as
$$\Sigma = \frac{1}{N}\tilde{X}^T\tilde{X}$$
3. Calculate the eigenvectors and eigenvalues of Σ
4. Pick d' eigenvectors corresponding to the largest eigen values and put them in the column of $A = [v_1, \ldots, v_{d'}]$
5. $X' = A^T X$

Fig. 4. The overall description of PCA.

1. Suppose two classes have mean values of μ_1 *and* μ_2 and covariances of Σ_1 and Σ_2
2. Estimate the between-groups scatter matrix S_B as
$$S_B = (\mu_1 - \mu_2)(\mu_1 - \mu_2)^T$$
3. Estimate the within-groups scatter matrix S_W as
$$S_W = S_1 + S_2$$
$$S_1 = \sum_{x_i \varepsilon C_1} (x_i - \mu_1)(x_i - \mu_1)^T$$
$$S_2 = \sum_{x_i \varepsilon C_2} (x_i - \mu_2)(x_i - \mu_2)^T$$
4. Consider A is the eigenvector corresponding to the largest eigenvalue of $S_W^{-1}S_B$
5. $X' = A^T X$

Fig. 5. The overall description of LDA feature selection.

Consider A is the eigenvector corresponding to the largest eigenvalue of $S_W^{-1}S_B$ $X^{'} = A^T X$

3.1.5. Genetic algorithm

Genetic Algorithm (GA) [23] is a stochastic search algorithm which imitates the mechanics of natural selection and natural genetics (natural competition between individuals for limited natural sources). The population obtains more natural resources by selecting and reproducing superior individuals. GA can simulate this process and calculate proper weights of risk factors in shunt infection dataset.

The overall description of GA is presented in Fig. 6. The coding of the individuals must first be defined. Then, an initial population of individuals is randomly created. Next, a set of operators is applied to generate successive populations, which hopefully evolve and improve through the time. The main GA operators are reproduction, crossover, and mutation. Reproduction is a process based on fitness function which identifies how good an individual is. Thus, individuals with higher fitness values have higher probability of contributing to the next generation. Crossover is a genetic operator used to combine the genetic information of two selected parents for generating new offsprings. Combining elements from two parents hopefully improves fitness values. Mutation is the random alteration of the elements of new offsprings with a small probability. In fact, mutation is a process of random walk through the search space to ensure that the important information contained within individuals may not be lost prematurely.

1. Generate the initial population randomly.
2. Repeat
   - Evaluate each individual by fitness function.
   - Select parents based on their fitness.
   - Generate children from the selected parents by crossover operator.
   - Mutate the new children by mutation operator.
   - Replace the new generation with the old generation.
3. Until termination criteria is met or a fixed number of generations have been accomplished.

In this study, the continuous version of GA was used for the feature selection task based on which each individual contained the weights associated with the aforementioned factors (length of each individual = 11). Each individual was evaluated based on its ability for the shunt infection prediction task. To determine the fitness value of each individual, factors with low weights were ignored, and the remaining weighted factors were entered to the classifier. The prediction accuracy of the classifier identifies how good an individual is. Table 2 shows the value of GA parameters.

1. Generate the initial population randomly.
2. Repeat
   - Evaluate each individual by fitness function.
   - Select parents based on their fitness.
   - Generate children from the selected parents by crossover operator.
   - Mutate the new children by mutation operator.
   - Replace the new generation with the old generation.
3. Until termination criteria is met or a fixed number of generations have been accomplished.

Fig. 6. The overall description of GA algorithm.

Table 2
Value of the GA parameters.

Parameter		Parameter	
Population size	50	Maximum iterations	100
Selection	Stochastic Uniform	Crossover	Constraint dependent
Crossover rate	0.8	Mutation	Constraint dependent
Mutation rate	0.05		

3.2. Classification

It is expected that the risk factors selected by the schemes described in the previous section have a good performance in comparison with all risk factors. To test this, some widely-used classifiers such as k-Nearest Neighbor (kNN), Support Vector Machine (SVM), random forest, Adaboost and Naive Bayes were chosen as candidate classifiers.

3.2.1. K-nearest neighbors

kNN [24] is a simple classifier which is very popular due to its simplicity of interpretation and can be implemented by parallel processing methods. A new sample is classified by a majority vote of its neighbors, assigned to the most common class among its k nearest neighbors.

3.2.2. Support Vector Machine

SVM [25] is naturally a binary classifier which performs classification tasks by maximizing the margin between the classes simultaneously while minimizing the empirical error simultaneously. Those training samples lying within the margin are considered as support vectors and determine the slope of the linear boundary with maximum margin in the feature space. Fig. 7 shows how the SVM classifier classifies its input samples.

1. Assume samples data $(x_i,\ i = 1, \ldots, N)$ with labels $y_i = \pm 1$.
2. Transform the data from the initial space R^n by a non-linear transformation (ø) to R^m space with more dimension as follows

$$x_i \ \text{ø}\ (x_i)$$

3. Compute the optimal linear boundary (b and W values) with the following conditions

$$W.X + b = 0$$

$$\text{subject to.} \begin{cases} y_i(wx_i + b) = 1 \ \textit{If } x_i \textit{ is not a support vector} \\ y_i(wx_i + b) > 1 \ \textit{If } x_i \textit{ is a support vector} \end{cases}$$

where x is a point on the decision boundary, W is an n-dimensional vector in perpendicular to the decision boundary, and b/|| w || shows the source distance to the decision boundary.

3.2.3. Random forest

Random forest [26] is an ensemble learning algorithm that builds multiple Decision Trees (DTs) and considers them together to get a more accurate and stable prediction. The results of several DTs can be combined by averaging or taking the majority vote. Each DT algorithm works by choosing the finest feature (in a random subset of predictors) to divide the data and expand the leaf nodes of the tree until the ending condition is met. In this study, the number of DTs in the random forest is set to 10 through cross validation.

3.2.4. Adaboost

Adaboost [27] is a set of weak classifiers, which are structured in parallel and learned sequentially. Each weak learner tries to learn a distribution of the data biased to the misclassified samples of the former learner. The idea of Adaboost is that the proper combination of weak learners can make a strong learner, where each learner compensates for the deficiency of the former by boosting the weights of the misclassified samples. In this study, DT was chosen as the weak learner, and the number of weak learners was set to 10 through cross validation. Fig. 8 shows the Adaboost algorithm.

1. Assume samples data $(x_i,\ i = 1, \ldots, N)$ with labels $y_i = \pm 1$.
2. Initialize all weights are set equally. In each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set.
3. Find a weak classifier $h_t : X \rightarrow Y$ that minimizes the prediction error ε_t. Then, α_t, the weight (importance) of h_t classifier, is updated as follows

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

4. The final classifier is a weighted majority vote of the T weak classifiers

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

3.2.5. Naive Bayes

Naive Bayes is a statistical classification technique based on the Bayes Theorem, which assumes that all features are mutually independent and conditional on the category. Under this assumption, Naive Bayes classifier consider a decision rule as follow

$$\hat{y} = argmax_{k\in\{1,2\}}\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k) \tag{1}$$

1 Calculate the prior probability $p(C_k)$ for given class labels
2. Calculate the posterior probability based on Bayes theorem as follows:

$$\hat{y} = argmax_{k\in\{1,2\}}\ p(C_k) \prod_{i=1}^{n} p(x_i|C_k)$$

3. Select the class with a higher probability. Therefore, the given input is assigned to the class with a higher probability.

Fig. 9The overall description of the Naive Bayes algorithm.

1. Assume samples data $(x_i, i = 1, \ldots, N)$ with labels $y_i = \pm 1$.
2. Transform the data from the initial space R^n by a non-linear transformation (ø) to R^m space with more dimension as follows

$$x_i \rightarrow \text{ø}(x_i)$$

3. Compute the optimal linear boundary (b and W values) with the following conditions

$$W.X + b = 0$$

$$\text{subject to} \begin{cases} y_i(wx_i + b) = 1 \ \text{If } x_i \text{ is not a support vector} \\ y_i(wx_i + b) > 1 \ \text{If } x_i \text{ is a support vector} \end{cases}$$

where x is a point on the decision boundary, W is an n-dimensional vector in perpendicular to the decision boundary, and b / || w || shows the source distance to the decision boundary.

Fig. 7. The overall description of the SVM algorithm.

1. Assume samples data $(x_i, i = 1, \ldots, N)$ with labels $y_i = \pm 1$.
2. Initialize all weights are set equally. In each round, the weights of incorrectly classified examples are increased so that the weak learner is forced to focus on the hard examples in the training set.
3. Find a weak classifier $h_t: X \rightarrow Y$ that minimizes the prediction error ε_t. Then, α_t, the weight (importance) of h_t classifier, is updated as follows

$$\alpha_t = \frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right)$$

4. The final classifier is a weighted majority vote of the T weak classifiers

$$H(x) = sign\left(\sum_{t=1}^{T} \alpha_t h_t(x)\right)$$

Fig. 8. The overall description of the Adaboost algorithm.

1. Calculate the prior probability $p(C_k)$ for given class labels
2. Calculate the posterior probability based on Bayes theorem as follows:

$$\hat{y} = argmax_{k\in\{1,2\}}\, p(C_k)\prod_{i=1}^{n} p(x_i|C_k)$$

3. Select the class with a higher probability. Therefore, the given input is assigned to the class with a higher probability.

Fig. 9. Displays the Naive Bayes algorithm.

4. Experimental results and discussion

In this study, 148 hydrocephalus patients including 68 with the history of shunt infection and 80 without any shunt infection (controls) were analyzed. In the first stage, two patients with missing features were removed from the dataset resulting in 146 samples (66 patients with hydrocephalus and 80 control cases). Then, well-known classifiers were applied to the dataset for the prediction task. The data set was divided into test and validation set. Test set contain 29 samples and validation set contain 117 samples. In the validation set, 5-fold cross validation is executed. Through this cross validation, the best parameters for each classifier are selected and finally the classifiers with the best parameters are applied to the test set and the overall performance is determined by taking an average over the final test samples. In the second stage, to compare the discriminative information of different factors, the prediction rate of each factor is reported in Table 3. As shown, the best results are related to number of shun revisions, birth weight, prematurity, and age of the first shunt procedure, respectively.

In the third stage, all factors were considered to compare the decision-making performance for the prediction task. Table 4 indicates the mean and Std of accuracy, sensitivity, and specificity of different classifiers. As shown, the prediction accuracy ranges from 63% to 75%, and the Adaboost and Random forest have the best results in comparison with other classifiers.

In the fourth stage, five state-of-art feature selection methods were applied to determine the most discriminative features (risk factor) for predicting shunt infection. Tables 5–7 shows the results of the aforementioned algorithm for the prediction task, indicating that the feature selection by SFS and GA enhances the performance of all classifiers. Moreover, the standard deviation of the results decreases for the classifiers. Thus, the robustness of the results increases while using feature selection. The prediction accuracy is 64–75% and 62–81% in SFS and GA approaches, respectively. The best feature selection results belong to GA which selects prematurity history, intraventricular hemorrhage history, age of the first shunt procedure, number of shunt revisions, brain tumor induced hydrocephalus, birth weight, and coinfection. The results of the present study are in line with Habibi et al. [5] which found that shunt infection risk factors are shunt revision history, low birth weight, prematurity history, the age of the first shunt procedure, intraventricular hemorrhage history, myelomeningocele history, and coinfection.

Fig. 10 shows the weights related to the selected factors by GA. These weights are assigned based on discriminative information included in each risk factor (the higher weight leads to more discriminative factor). As displayed in Fig. 10, prematurity and coinfection have the highest and lowest weights, respectively. Now, the clinical importance of the features which are selected by GA is explained. These factors are grouped into significant factors and non-significant factors as follows:

1) *Prematurity:* Several studies demonstrated prematurity as an independent risk factor for shunt infection [28–30]. Premature neonates had undeveloped immune system and skin barrier along with more pathogen bacterial flora of skin, which can lead to higher occurrence of shunt infection in this group. The findings showed that GA gave the highest weight to this factor. Therefore, it can be more appropriate for the prediction task. Additionally, for more comparison, the difference between the prematurity of the two groups was evaluated by using *t*-test. The confidence level of $p < 0.05$ was considered as statistically significant. Based on the results, a significant difference was observed between the two groups in terms of prematurity ($p < 0.05$). Hence, neurosurgeons should postpone shunt insertion as much as possible for the patients.
2) *Number of shunt revisions:* More than 50% of the patients with a ventricular shunt require at least one revision surgery [31]. Thus, each surgery potentially produces an opportunity for new organisms to infect the shunt device and the operation site. Several studies reported an independent relationship between the number of shunt revisions and shunt infection risk [32–34]. In the present study, this factor was also selected based on the ability for the predication task. The results demonstrated a significant difference between the two

Table 3
The prediction rate of different features with different classifiers.

	KNN Mean ± Std	SVM Mean ± Std	Randomforest Mean ± Std	Adaboost\ Mean ± Std	Naive Bayes Mean ± Std
Birth weight	67.59 ± 9.93	55.86 ± 7.48	64.83 ± 12.58	61.38 ± 11.28	59.31 ± 6.17
Prematurity	62.07 ± 7.31	62.76 ± 12.04	62.07 ± 5.97	62.07 ± 9.44	62.07 ± 10.05
Trauma	40.69 ± 6.63	55.17 ± 14.83	51.03 ± 8.93	53.79 ± 10.23	48.28 ± 3.45
Meningitis	42.76 ± 3.93	54.48 ± 7.86	42.76 ± 13.92	52.41 ± 7.48	48.97 ± 10.74
Tumor	51.72 ± 5.45	54.48 ± 4.50	54.48 ± 8.23	51.72 ± 8.79	42.07 ± 4.50
IVH	45.52 ± 7.48	53.10 ± 8.30	53.79 ± 11.59	57.24 ± 3.93	53.10 ± 6.72
Icter	55.86 ± 4.50	58.62 ± 6.90	57.93 ± 7.48	58.62 ± 8.79	58.62 ± 6.45
Coinfection	44.14 ± 4.50	55.17 ± 7.31	56.55 ± 11.07	56.55 ± 8.99	56.55 ± 11.33
MMC	55.86 ± 7.48	48.97 ± 8.93	50.34 ± 3.93	33.79 ± 22.66	55.86 ± 3.78
Age at first shunt	64.83 ± 7.86	43.45 ± 10.23	64.14 ± 11.59	63.45 ± 9.32	42.76 ± 10.52
No of Shunt Revision	72.41 ± 8.09	71.72 ± 7.48	66.21 ± 11.79	72.41 ± 8.45	62.76 ± 8.93

Table 4
The prediction accuracy of different classifiers with all features.

Classifier	Specificity Mean ± Std	Sensitivity Mean ± Std	Prediction accuracy Mean ± Std
kNN	74.17 ± 14.13	62.11 ± 13.95	68.97 ± 5.45
SVM	68.97 ± 23.09	59.91 ± 14.72	63.45 ± 9.93
Random forest	73.38 ± 10.32	75.78 ± 13.89	73.79 ± 6.72
Adaboost	79.12 ± 7.40	71.94 ± 16.33	75.42 ± 8.41
Naive Bayes	83.56 ± 13.89	48.91 ± 3.09	66.90 ± 6.72

Table 5
The specificity of different classifiers with selected features.

	SFS	LASSO	PCA	LDA	GA
kNN	72.18 ± 9.01	71.83 ± 15.16	68.88 ± 3.99	100.00 ± 00.00	78.36 ± 8.22
SVM	83.12 ± 10.24	81.73 ± 18.02	80.57 ± 12.15	81.23 ± 31.79	65.28 ± 25.54
Random forest	71.96 ± 8.98	69.70 ± 10.98	80.46 ± 8.73	68.88 ± 6.80	82.22 ± 12.25
Adaboost	74.96 ± 16.66	72.69 ± 5.72	74.07 ± 13.79	80.33 ± 12.52	79.53 ± 13.70
Naive Bayes	87.45 ± 11.46	82.78 ± 14.47	89.72 ± 5.34	85.48 ± 8.24	81.01 ± 6.38

Table 6
The sensitivity of different classifiers with selected features.

	SFS	LASSO	PCA	LDA	GA
kNN	75.25 ± 16.65	70.52 ± 15.30	59.89 ± 16.92	00.00 ± 00.00	85.91 ± 5.18
SVM	67.66 ± 15.41	61.42 ± 15.67	56.47 ± 11.02	21.73 ± 29.14	58.19 ± 9.81
Random forest	69.44 ± 19.42	67.22 ± 13.10	63.07 ± 12.60	56.19 ± 10.49	75.08 ± 8.23
Adaboost	78.49 ± 12.66	69.31 ± 14.18	79.24 ± 5.73	51.52 ± 17.78	75.54 ± 15.79
Naive Bayes	37.33 ± 12.78	43.91 ± 16.38	42.57 ± 16.33	27.41 ± 16.50	55.14 ± 13.85

Table 7
The accuracy of different classifiers with selected features.

	SFS	LASSO	PCA	LDA	GA
kNN	72.41 ± 4.88	71.03 ± 8.99	64.83 ± 7.48	54.48 ± 5.67	81.38 ± 5.23
SVM	75.86 ± 10.90	72.41 ± 8.09	69.66 ± 6.63	53.10 ± 9.63	62.07 ± 13.58
Random forest	70.34 ± 9.32	68.97 ± 11.17	70.34 ± 5.23	62.76 ± 2.89	77.93 ± 8.30
Adaboost	75.86 ± 9.44	70.34 ± 9.93	76.55 ± 8.93	68.28 ± 3.78	76.55 ± 4.50
Naive Bayes	64.14 ± 5.23	65.52 ± 10.05	67.59 ± 9.63	59.31 ± 8.23	70.34 ± 3.08

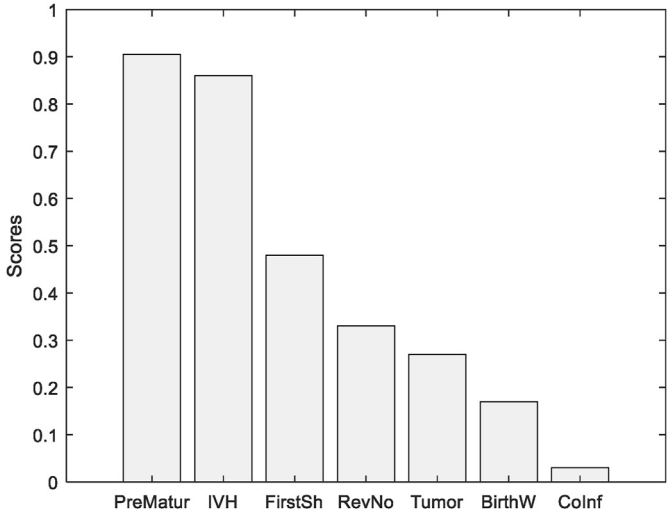

Fig. 10. The weight of the selected risk factors by GA.

groups in terms of the number of shunt revisions ($p < 0.05$). It seems that further focus on optimized revision procedure can decrease shunt infection risks.

3) *Low birth weight:* Based on the results of the previous studies, low birth weight infants had higher rates of shunt infection [32,35,36] due to their immature immune system. As shown in Fig. 10, GA ranks this factor as the sixth. A significant difference was observed between the two groups in terms of birth weights ($p < 0.05$). Therefore, any possible delays in the shunt placement for increasing the neonatal weight can lead to lower shunt infection occurrence.

Non-significant factors

4) *Intra-Ventricular Hemorrhage (IVH):* Neonates with post-hemorrhagic hydrocephalus have a higher risk of shunt infection [30], but it is controversial if it is an independent variable or due to other coexisting factors like prematurity and low birth weight. As shown in Fig. 10, this factor ranks the second after prematurity history. However, the results indicated that there was no significant difference between the two groups in terms of IVH ($p > 0.05$).
5) *Age of the first shunt procedure*: Several studies showed a considerable higher risk of shunt infection in pediatric patients compared to adult population [37–39] and a significant relationship between age (less than 6 months) and occurrence of shunt infection [28,32,40]. As Fig. 10 shows, this factor ranks the third after the history of prematurity and IVH, which may be related to the poorly developed immune system of the neonates and immaturity of their skin barrier. Nevertheless, there was no significant difference between the two groups in terms of the age of the first shunt procedure ($p > 0.05$).
6) *Brain tumor:* Other neurosurgical operations can expose CSF with bacterial pathogens and potentially increase shunt infection risk. However, it was not considered as a strong risk factor for shunt infection in the literature. In addition, neurosurgeons usually manage tumoral induced hydrocephalus with other interventions except shunt insertion. In this study, GA selected this factor based on the predication rate of shunt infection for the two groups. However, the results indicated no significant difference between the two groups in terms of brain tumor ($p > 0.05$).
7) *Coinfection:* Blood borne infections such as appendicitis and bowel perforation may play an important role in late shunt infections [38,41]. However, they are not considered as a prevalent cause of shunt infection in general. As displayed in Fig. 10, this factor has the lowest weights among the selected factors. Furthermore, there was no significant difference between the two groups in terms of coinfection ($p > 0.05$).

5. Conclusion

To clarify the complexity of the infection management, this study analyzed the classification techniques and feature selection/weighting schemes to select or give a proper weight to the features for predicting the surgical infection risk. Based on the results, three risk factors including prematurity, number of shunt revisions, and low birth weight constituted substantial parts of the predictive performance. Hence, neurosurgeons should postpone shunt insertion as much as possible in premature and low birth weight patients. Further, given the number of shunt revisions highlighted in the prediction task, it is suggested that neurosurgeons carefully perform shunt insertion based on standard protocols for the patients.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Cinalli G, Maixner WJ, Sainte-Rose C. Pediatric hydrocephalus. Springer-Verlag. 2005.
[2] Gutierrez-Murgas Y, Snowden JN. Ventricular shunt infections: immunopathogenesis and clinical management. J Neuroimmunol 2014;276:1–8. 0.
[3] Kestle JR, Holubkov R, Cochrane D, Kulkarni AV, et al. A new hydrocephalus clinical research network protocol to reduce cerebrospinal fluid shunt infection. J Neurosurg Pediatr 2016;17(4):391–6.
[4] Lee JK, Seok JY, Lee JH, Choi EH, et al. Incidence and risk factors of ventriculoperitoneal shunt infections in children: a study of 333 consecutive shunts in 6 years. J Kor Med Sci 2012;27(12):1563–8.
[5] Habibi Z, Ertiaei A, Nikdad MS, Mirmohseni AS, et al. Predicting ventriculoperitoneal shunt infection in children with hydrocephalus using artificial neural network. Child's Nerv Syst 2016;32:2143–51.
[6] Luz CF, Vollmer M, Decruyenaere J, Nijsten MW, et al. Machine learning in infection management using routine electronic health records: tools, techniques, and reporting of future technologies. Clin Microbiol Infect 2020;26(10):1291–9.

[7] Jothia N, Rashid NA, Husain W. Data mining in healthcare – a review. Procedia Comput Sci 2015;72:306–13.
[8] Shahid N, Rappon T, Berta W. Applications of artificial neural networks in health care organizational decision-making: a scoping review. PloS One 2019;14(2): e0212356.
[9] Raghavendra U, Acharya UR, Adeli H. Artificial intelligence techniques for automated diagnosis of neurological disorders. Eur Neurol 2019;82:41–64.
[10] Azimi P, Mohammadi HR. Predicting endoscopic third ventriculostomy success in childhood hydrocephalus: an artificial neural network analysis. J Neurosurg Pediatr 2014;13(4):426–32.
[11] Azimi P, Mohammadi HR, Benzel EC, Shahzadi S, Azhari S, Montazeri A. Artificial neural networks in neurosurgery. J Neurol Neurosurg Psychiatr 2015;86(3):251–6.
[12] Seattle children's hospital. https://www.seattlechildrens.org/conditions/hydrocephalus.
[13] Remeseiro B, Bolon-Canedo V. A review of feature selection methods in medical applications. Comput Biol Med 2019;112:103375.
[14] Peng Y, Wu Z, Jiang J. A novel feature selection approach for biomedical data classification. J Biomed Inf 2010;43(1):15–23.
[15] Sabeti M, Boostani R, Katebi SD, Price GW. Selection of relevant features for EEG signal classification of schizophrenic patients. Biomed Signal Process Contr 2007;2: 122–34.
[16] Holzinger A. Machine learning for health informatics: state-of-the-art and future challenges. Springer. 2016.
[17] Sabeti M, Katebi SD, Boostani R, Price GW. A new approach for EEG signal classification of schizophrenic and control participants. Expert Syst Appl 2011; 38(3):2063–71.
[18] Alimardani F, Boostani R, Azadehdel M, Ghanizadeh A, Rastegar K. Presenting a new search strategy to select synchronization values for classifying bipolar mood disorders from schizophrenic patients. Eng Appl Artif Intell 2013;26(2):913–23.
[19] Sabeti M, Katebi SD, Boostani R. Entropy and complexity measures for EEG signal classification of schizophrenic and control participants. Artif Intell Med 2009;47: 263–74.
[20] Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc B 1996;58(1):267–88.
[21] Naik GR. Advances in principal component analysis, research and development. Singapore: Springer; 2018.
[22] Izenman AJ. Linear discriminant analysis. In: Modern multivariate statistical techniques. New York: Springer; 2013.
[23] Affenzeller M, Wagner S, Winkler S, Beham A. Genetic algorithms and genetic programming, modern concepts and practical applications. Chapman and Hall/ CRC. 2018.
[24] Biau G, Devroy L. Lectures on the nearest neighbor method, Springer. 2015.
[25] Keshavarz A, Ghasamiyan H. A fast algorithm based on support vector machine for classifying super-spectral images by spatial correlation. J Electr Comput Eng Iran 2005;3(1):250–63.
[26] Lior R, Maimon O. Data mining with decision trees: theory and applications. World Scientific Inc. 2008.
[27] Misra S, Li H. Chapter 9 - noninvasive fracture characterization based on the classification of sonic wave travel times. Mach Learn Subsurf Charact 2020:243–87.
[28] Kulkarni AV, Drake JM, Lamberti-Pasculli M. Cerebrospinal fluid shunt infection: a prospective study of risk factors. J Neurosurg 2001:95–201.
[29] Moussa WM, Mohamed MA. Efficacy of postoperative antibiotic injection in and around ventriculoperitoneal shunt in reduction of shunt infection: a randomized controlled trial. Clin Neurol Neurosurg 2016;143:144–9.
[30] Spader HS, Hertzler DA, Kestle JR, Riva-Cambrin J. Risk factors for infection and the effect of an institutional shunt protocol on the incidence of ventricular access device infections in preterm infants. J Neurosurg Pediatr 2015;15(2):156–60.
[31] Wells DL, Allen JM. Ventriculoperitoneal shunt infections in adult patients. AACN Adv Crit Care 2013;24(1):6–12.
[32] Simon TD, Whitlock KB, Riva-Cambrin J, Kestle JR, Rosenfeld M, Dean JM, et al. Revision surgeries are associated with significant increased risk of subsequent cerebrospinal fluid shunt infection. Pediatr Infect Dis J 2012;31(6):551–6.
[33] Rogers EA, KimiaA Madsen JR, Nigrovic LE, Neuman MI. Predictors of ventricular shunt infection among children presenting to a pediatric emergency department. Pediatr Emerg Care 2012;28(5):405–9.
[34] Simon TD, Butler J, Whitlock KB, Browd SR, Holubkov R, Kestle JR. Risk factors for first cerebrospinal fluid shunt infection: findings from a multi-center prospective cohort study. J Pediatr 2014;164(6):1462–8.
[35] Bruinsma N, Stobberingh EE, Herpers MJ, Vles JS, Weber BJ, Gavilanes DA. Subcutaneous ventricular catheter reservoir and ventriculoperitoneal drain-related infections in preterm infants and young children. Clin Microbiol Infect 2000;6(4): 202–6.
[36] Dallacasa P, Dappozzo A, Galassi E, Sandri F, Cocchi G, Masi M. Cerebrospinal fluid shunt infections in infants. Child's Nerv Syst 1995;11(11):643–8.
[37] Choux M, Genitori L, Lang D, Lena G. Shunt implantation: reducing the incidence of shunt infection. J Neurosurg 1992;77(6):875–80.
[38] Reddy GK, Bollam P, Caldito G. Ventriculoperitoneal shunt surgery and the risk of shunt infection in patients with hydrocephalus: long-term single institution experience. World Neurosurg 2012;78(1–2):155–63.
[39] Vinchon M, Dhellemmes P. Cerebrospinal fluid shunt infection: risk factors and long-term follow-up. Child's Nerv Syst 2006;22(7):692–7.
[40] Braga MH, Carvalho GT, Brandao RA, Lima FB, Costa BS. Early shunt complications in 46 children with hydrocephalus. Arq Neuropsiquiatria 2009;67(2A):273–7.
[41] Vinchon M, Lemaitre MP, Vallee L, Dhellemmes P. Late shunt infection: incidence, pathogenesis, and therapeutic implications. Neuropediatrics 2002;33(4):169–73.