Available online at sjuoz.uoz.edu.krd

Science Journal of University of Zakho

Vol. 8, No. 3, pp. 95– 111, September-2020

p-ISSN: 2663-628X
e-ISSN: 2663-6298

A FULLY BAYESIAN LOGISTIC REGRESSION MODEL FOR CLASSIFICATION OF ZADA DIABETES DATASET

Masoud M. Hassan [a]

[a] Dept. of Computer Science, Faculty of Science, University of Zakho, Kurdistan Region, Iraq – (Masoud.hassan@uoz.edu.krd)

Received: Mar., 2020 / Accepted: Jun., 2020 / Published: Sep.,2020 ***https://doi.org/10.25271/sjuoz.2020.8.3.707***

ABSTRACT:
Classification of diabetes data with existing data mining and machine learning algorithms is challenging and the predictions are not always accurate. We aim to build a model that effectively addresses these challenges (misclassification) and can accurately diagnose and classify diabetes. In this study, we investigated the use of Bayesian Logistic Regression (BLR) for mining such data to diagnose and classify various diabetes conditions. This approach is fully Bayesian suited for automating Markov Chain Monte Carlo (MCMC) simulation. Using Bayesian methods in analysing medical data is useful because of the rich hierarchical models, uncertainty quantification, and prior information they provide. The analysis was done on a real medical dataset created for 909 patients in Zakho city with a binary class label and seven independent variables. Three different prior distributions (Gaussian, Laplace and Cauchy) were investigated for our proposed model implemented by MCMC. The performance and behaviour of the Bayesian approach were illustrated and compared with the traditional classification algorithms on this dataset using 10-fold cross-validation. Experimental results show overall that classification under BLR with informative Gaussian priors performed better in terms of various accuracy metrics. It provides an accuracy of 92.53%, a recall of 94.85%, a precision of 91.42% and an F1 score of 93.11%. Experimental results suggest that it is worthwhile to explore the application of BLR to predictive modelling tasks in medical studies using informative prior distributions.

KEYWORDS: Diabetes, Bayesian Logistic Regression, Markov Chain Monte Carlo, Classification, Informative Priors.

1. INTRODUCTION

Diabetes mellitus is one of the most widespread and challenging diseases worldwide. People with diabetes usually suffer from high blood sugar. The reduction of insulin production and losing its efficiency in metabolism are the two major causes of this disease (Hassan & Amiri, 2019). Diabetics could have critical complications such as blindness, heart attack, kidney failure, and strokes. The main symptoms of diabetes are intensified thirst, hunger, and frequent urination.

Logistic regression is a common statistical method that has been extensively used for classification problems and predicting binary responses in medical studies (Maxime Vono, 2018). Over the last decade, this model has been successfully used in many fields, including business, medicine (Chang et al., 2018) and social sciences (Wang et al., 2010). Interpretation of the estimated logistic regression coefficients is an important task to take the right decision. However, the standard Logistic Regression (LR) model cannot always provide such proper interpretation under this consideration. To cope with this issue, using Bayesian models will help provide a better understanding of the estimated parameters of the model (Joseph, 2016). Bayesian inference takes into account the combination of prior information about the unknown parameters with the likelihood of data to derive probabilistic interpretation of the model via estimation of the posterior distributions of the model coefficients (Clark et al., 2007).

In this paper, we investigate the use of Bayesian logistic regression (BLR) models with the use of Markov chain Monte Carlo (MCMC) methods to classify diabetics. The BLR model combines prior knowledge with the LR model in a Bayesian framework (Hassan et al., 2019). The main advantage of using Bayesian models is the combination of prior knowledge with data so that one can include past information about unknown parameters and form prior distributions to estimate the posterior probabilities and learn about the true values of the parameters (Hassan, 2018). Consequently, when new records become available, the previous posterior distributions can be used as priors. In this paper, we used different informative and non-informative priors to evaluate the performance of the Bayesian analysis to classify diabetes patients.

This study is the first to use a Bayesian model for predicting diabetes for our new ZADA dataset based on a Bayesian approach. ZADA is a newly created dataset consists of 909 patients which were collected from approximately 7,000 medical records of patients in Zakho city/ Kurdistan Region of Iraq. The dataset consists of seven independents variables and one class label indicating whether the patient is diabetic or healthy. The fully Bayesian (MCMC) approach was used to automatically classify diabetics (sick or normal) with a Bayesian binomial logistic regression model applied to linear combinations of seven medical features. To the best of our knowledge, there have been only a few studies in the literature investigating fully Bayesian methods for classification problems and examining regression posteriors of unknown parameters based on informative priors. In our Bayesian modelling, we investigated the use of three different priors suggested in the literature: (1) informative Gaussian (Gelman et al., 2008); (2) a weakly informative Laplace (Li & Yao, 2018); (3) a weakly informative Cauchy (Ghosh et al., 2018a). Therefore, the performance of the Bayesian classification model under each prior used was evaluated by comparison of agreement between the true and predictive (model) classification of diabetes.

To fit any classification/regression model, we must take correlations among attributes into account when attempting to capture the distribution of the model outcome (binary classes) given all the independent attributes. However, when the number of samples is less than or equal to the number of attributes, the

classification model might overfit the data, which will capture the noise rather than the signal (Li & Yao, 2018). Hence, different priors for the Bayesian approach under study can be used in such cases to avoid overfitting problems.

The rest of this paper is organized as follows. In section 2, we reviewed some relevant works of using logistic regression models and Bayesian approaches in particular. In section 3, we presented the proposed mathematical approach of our BLR model. In Section 4, we presented the experiments conducted on the ZADA dataset and compared the results of BLR with other well-known traditional classifiers. Finally, we concluded our work and discussed future works in Section 5.

2. RELATED WORKS

Although Bayesian methods have been widely used in the literature for different statistical analysis, they are still in lack of development to be used in the machine learning domain for classification problems. In this section we review some relevant papers using logistic regression in the Bayesian framework.

Ghosh et al. (Ghosh et al., 2018a) used a weakly Cauchy prior for Bayesian Logistic Regression. They studied the presence of posterior summaries based on Cauchy priors. In their implementation, they developed a Gibbs sampling algorithm using Gamma data augmentation to draw samples from the posterior distributions based on different priors. Their empirical results showed that even when the mean of the posteriors was used for Cauchy priors, the posterior estimates of the model parameters might be very large.

Li et al. (Li & Yao, 2018) proposed a Bayesian logistic regression based on hyper-Lasso priors as a feature selection method for high-dimensional. They used a Hamiltonian Monte Carlo sampling algorithm to learn about logistic regression coefficients. Their experimental results on simulated and real microarray data showed higher performance of hyper-Lasso comparing to classical Lasso for feature selection.

Suleiman et al. (Suleiman et al., 2019) used BLR models to predict incorrect Diagnosis-Related Group DRG assignment. They used weakly informative priors in their investigation to find the likelihood of their DRG revision, hence comparing the Bayesian model estimates with classical maximum likelihood estimates. The experimental results of their comparative study showed that the use of Bayesian methods could improve model parameters stabilization and model's classification performance.

Octaviani et al. (Octaviani et al., 2019) and DuMouchel (DuMouchel, 2012) used multivariate Bayesian logistic regression for analysis of clinical study safety issues. They proposed a Bayesian logistic regression for ovarian cancer classification. The accuracy of their proposed mothed was around 70%.

Chang et al. (Chang et al., 2018) applied Bayesian Logistic Regression to predict breast cancer using the Wisconsin Diagnosis Breast Cancer (WDBC) data. They compared their Bayesian approach with other traditional algorithms: Decision Trees, Random Forest, Neural Network, Support Vector Machine (SVM), and Logistic Regression (LR). They concluded that BLR could provide better classification performance.

Spyroglou et al. (Spyroglou et al., 2018) used a Bayesian Logistic Regression approach for predicting asthma persistence in children. Due to multicollinearity exist in their explanatory features, they used Principal Component Analysis (PCA) combined with the Bayesian logistic regression to analyze and predict data of 147 asthmatic children. Their approach was implemented by (MCMC) algorithms. They concluded that the proposed method can accurately predict asthma with high accuracy and provides better knowledge about the influence and importance of each factor in predicting asthma persistence.

Holmes and Held (Holmes & Held, 2006) investigated the use of auxiliary variable methods for BLR, considering covariate set uncertainties. They proposed a fully automatic approach with no user-set parameters and no necessary for Metropolis-Hastings accept/reject steps. They concluded that their method is successful and provided a fast, effective automatic algorithm to make inference about model parameters.

Lin et al. (Lin et al., 2019) used a Bayesian hierarchical logistic regression model for analysing multiple informant family health history (FHH) of diabetes. They used informative priors for modelling the disease for integrating multiple informant FHH for risk prediction purposes. Their experimental results based on Bayesian modelling indicated that for diabetic patients, both disease history and health behaviour information are important based on their posterior distributions being studied.

Wang et al. (Wang et al., 2010) investigated the use of Bayesian logistic regression model for spoken language identification (LID) to address the overfitting issues in their model. Their approach was able to accurately classify the NIST LRE 2007 dataset with the issue of overfitting being controlled.

3. METHODOLOGY

3.1 Bayesian Logistic Regression for Classification

Logistic regression usually models the relationship between a binary response variable (in our case is diabetic or healthy) with the explanatory variables which can be continuous or categorical. The Bayesian inference of logistic regression is somehow similar to the Bayesian linear regression model. However, logistic regression is even simpler for there is no variance term to be estimated, and only the regression parameters will be estimated.

3.2 The Likelihood:

Suppose that the response variable $\boldsymbol{y}$ is binary (Diabetic =1 or Healthy=0) with respective probability p and $(1-p)$. The logistic regression can be defined as (Mary Gladence et al., 2015):

$$\log\left(\frac{p}{1-p}\right) = \beta_1 + \beta_2 x_1 + \beta_3 x_2 + \cdots + \beta_i x_i. \quad (1)$$

Where $\boldsymbol{X} = (x_1, x_2, \dots, x_i)$ is a vector of the independent explanatory variables (features), $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \dots, \beta_i)$ is a vector of the unknown regression parameters of the model, and $i = 1,2, \dots k$ is the number of features. Therefore, the predicted value of $\boldsymbol{y}$ can be formulated from Equation 2 as follows (Octaviani et al., 2019):

$$E(y) = P(y=1) = \frac{exp(\beta_1+\beta_2 x_1+\beta_3 x_2+\cdots+\beta_i x_i)}{1+exp(\beta_1+\beta_2 x_1+\beta_3 x_2+\cdots+\beta_i x_i)} \quad (2)$$

Where,

$$y = \begin{cases} 0; \text{ Normal} \\ 1; \text{ Diabetic} \end{cases}$$

Equation 2 can also be expressed as follows:

$$E(y) = \frac{\exp(\boldsymbol{\beta X})}{1+\exp(\boldsymbol{\beta X})} \quad (3)$$

Now, given some training samples $(y_j, x_{1j}, x_{2j}, \dots, x_{ij})$, where $j = 1,2, \dots n$ is the number of samples, and y_j are n independent realizations of a Bernoulli experiment with the probability of success $P(y_j = 1)$ given by Equation 2. Thus, the likelihood function of the training sample is defined as follows (Madigan et al., 2005):

$$L(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{X}) = \prod_{j=1}^{n} P_j^{y_j}(1-P_j)^{1-y_j} \quad (4)$$

The Equation 4 can be rewritten as follows (Huggins et al., 2016):

$$L(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{X}) = \prod_{j=1}^{n}\left[\left(\frac{\exp(\boldsymbol{\beta X})}{1+\exp(\boldsymbol{\beta X})}\right)^{y_j}\left(1-\frac{\exp(\boldsymbol{\beta X})}{1+\exp(\boldsymbol{\beta X})}\right)^{1-y_j}\right] \quad (5)$$

In the classical statistical inference, the vector of the logistic regression parameter $\boldsymbol{\beta}$ for the model above can be estimated using the Maximum Likelihood Estimation (MLE) method (Chang et al., 2018),

$$\sum_{j=1}^{n}\left[y_j \log(P_j) + (1-y_j)\log(1-P_j)\right] \quad (6)$$

However, in this paper, we will be use Bayesian methods to estimate the model parameters of this logistic regression model to be used for classification and prediction of ZADA diabetes dataset.

3.2.1 The Prior: In order to make a Bayesian inference for the unknown parameters $\beta_1, \beta_2, \ldots, \beta_i$, we have to identify a prior distribution for each model parameters. The key of any Bayesian inference is how to choose such prior probability distributions. In this paper, we aim to use three informative priors to investigate the classification problem in the Bayesian framework. The three priors are:

1- Gaussian prior distribution:

$$P(\beta_j|\mu_j,\sigma_j^2) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(\beta_j - \mu_j)^2\right\} \quad (7)$$

2- Laplace prior distribution:

$$P(\beta_j|\mu_j,\lambda_j) = \frac{\lambda_j}{2} exp\{-\lambda_j|\beta_j - \mu_j|\} \quad (8)$$

3- Cauchy prior distribution:

$$P(\beta_j|\mu_j,\gamma_j) = \frac{1}{\pi\gamma_j\left[1+\left(\frac{\beta_j-\mu_j}{\gamma_j}\right)^2\right]} \quad (9)$$

The values for hyper-parameters $\mu_j, \sigma_j^2, \lambda_j,$ and γ_j must be chosen in such a way that they give informative priors. In this paper, we chose different values for hyper-parameters under each prior used as given below and shown in Figure 1:

$$\beta_j \sim Normal(\mu_j, \sigma_j^2)$$

$$\beta_j \sim Laplace(\mu_j, \lambda_j)$$

$$\beta_j \sim Cauchy(\mu_j, \gamma_j)$$

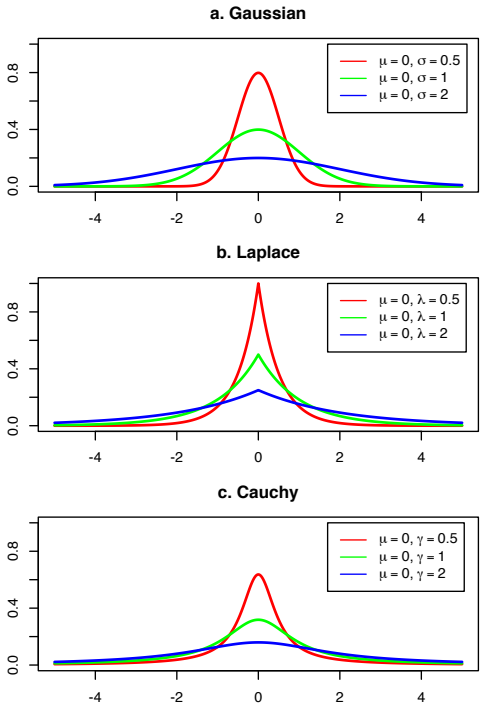

Figure 1. Prior distributions with different location and scale hyperparameters for Gaussian (**a**), Laplace (**b**), and Cauchy (**c**) to be used for Bayesian logistic model.

3.2.2 The Posterior: Bayesian inference allows a combination of prior beliefs about model parameters with the likelihood of data. In practice, Bayesian inference defines the uncertainty of each parameter as a statistical probability distribution called prior probability distributions. It hence derives posterior probability distributions by multiplying the prior distribution by the full likelihood function. Therefore, the posterior distribution of the unknown parameters β_j for the Bayesian logistic regression with Gaussian prior is given by (Suleiman et al., 2019):

$$\begin{aligned} \boldsymbol{Posterior} &\propto Prior \times Likelihood \\ P(\beta_j|\boldsymbol{y},\boldsymbol{X}) &\propto L(\boldsymbol{y}|\boldsymbol{\beta},\boldsymbol{X}) \times P(\beta_j) \\ &\propto \prod_{j=1}^{n}\left[\left(\frac{\exp(\boldsymbol{\beta X})}{1+\exp(\boldsymbol{\beta X})}\right)^{y_j}\left(1-\frac{\exp(\boldsymbol{\beta X})}{1+\exp(\boldsymbol{\beta X})}\right)^{1-y_j}\right] \\ &\quad \times \left[\prod_{j=1}^{n}\frac{1}{\sqrt{2\pi\sigma^2}} exp\left\{-\frac{1}{2\sigma^2}(\beta_j - \mu_j)^2\right\}\right] \end{aligned} \quad (10)$$

Practically, we cannot evaluate this posterior distribution, Equation 10, analytically. Hence, we should use a Markov chain Monte Carlo (MCMC) simulations which allow numerical sampling from underlying posterior distributions. The reader is referred to (Kass et al., 1997) for general information about MCMC algorithms and to (M. M. Hassan et al., 2019) and (Masoud Muhammed Hassan, 2018) for examples of implementing MCMC for multilevel statistical models. In this paper, we will use the Metropolis-Hasting with Gibbs sampler to estimate the marginal posterior distributions for our model parameters. The expected value of the posterior distribution of parameters β_j will be considered as regression coefficients of the Bayesian logistic model. One can also calculate the 95% highest posterior density (HPD) confidence intervals of the parameters from the estimated marginal distributions.

3.3 Bayesian Logistic Regression for Classification

Having successfully defined the BLR model, described in Section 3.1 above, and after estimating the parameters of the model, the performance of the proposed classification model can be evaluated. There are various performance metrics for multi-

level classifiers exist in machine learning literature (Suleiman et al., 2019). We used four commonly used performance metrics in our model evaluation; they are accuracy, precision, recall and F_1 (Chang et al., 2018). To calculate these evaluation metrics, we first need to calculate the confusion matrix, given in Table 1. Then the performance of the proposed model is compared with five well-known algorithms namely DT, K-NN, SVM, LR, and Naïve Bayes.

Table 1. Confusion matrix

	True Y	
Predicted Y	**0**	**1**
0	$M_{0,0}$ (TP)	$M_{0,1}$ (FP)
1	$M_{1,0}$ (FN)	$M_{1,1}$ (TN)

From the confusion matrix, all the evaluation metrics are calculated as follows (Chang et al., 2018):

$$Accuracy = \frac{\sum_{j=0}^{1} M_{j,j}}{\sum_{i=0}^{1}\sum_{j=0}^{1} M_{i,j}} \quad (11)$$

$$Precision = \frac{M_{0,0}}{\sum_{j=0}^{1} M_{0,j}} \quad (12)$$

$$Recall = \frac{M_{0,0}}{\sum_{j=0}^{1} M_{j,0}} \quad (13)$$

$$F_1 = \frac{2\times Precision \times Recall}{(Precision+Recall)} \quad (14)$$

Where $j = 0,1$ is the number of classes. The above metrics can provide the performance of the classification model, where TN, FP, FN, TP respectively denote the number of true negatives, false positive, false negative and true positives.

4. EXPERIMENTAL RESULTS

In order to assess the performance of the proposed method, the BLR model, defined in Section 3, was applied for classifying diabetes patients. To check the efficiency of the proposed BLR models, we compared the performance of the Bayesian classifiers with five other classical models: Decision Tree (DT), Support Vector Machine (SVM), K-Nearest Neighbor (K-NN), standard Logistic Regression (LR) and Naïve Bayes. The BLR models were experimentally evaluated using a newly created dataset, called ZADA. All the results were obtained using R language (R Development Core Team, 2011) and RStudio (Rstudio Team, 2019) with some particular tglm (Ghosh et al., 2018b) and rjags (Plummer, 2016) packages, on a computer equipped with 1.7 GHz Dual-Core Intel Core i7 processor with 8.0 GB of RAM.

All Bayesian experiments conducted in this paper were implemented using Metropolis-Hasting within the Gibs Sampler method. Each experiment was based on three independent runs with 20,000 iterations for each, and the convergence of the MCMC samplers was checked before reporting the results for the posterior inference. The number of burn-in samples was set to 2000. For implementing and validating our proposed approach, the ZADA dataset was first split into an 80% training set and a 20% testing set. The trainsets were used for model training and the test set was held out for model validation. For the classical classifiers, we also used a 10-fold cross-validation method to guarantee the randomness of the experiments as well as to avoid any modelling issues with underfitting or overfitting.

4.1 Dataset

The dataset used here is blood analysis of fasting sugar from Shaker laboratory in Zakho city, Kurdistan Region of Iraq. This dataset has not been used in any data mining application yet, and this is the first analysis of these data. This dataset contains different medical features for nearly 7,000 patients. After pre-processing (cleaning, integration, and reduction), we only selected the features which affect diabetes, and hence we came out with a new dataset, called ZADA, for diabetes only. This dataset has 909 records on seven features including the binary response feature "Class" which takes values 0 (healthy) and 1 (diabetic). The general characteristics of the ZADA diabetes dataset are summed up in Table 2.

Table 2. General characteristics of ZADA dataset

Attribute Name	Attribute Description	Min	Max	Mean
Age	Age of patients	20	86	48.01
Cholesterol	Test of Cholesterol	110	340	200.56
L_HDL	High-density Lipoprotein	23	65	42.97
L_LDL	Low-density Lipoprotein	36.8	266.2	124.87
L_VLDL	Very Low-Density Lipoprotein	8.6	80	32.73
Uric Acid	Test of Uric Acid	2.22	10.2	5.72
Class	1= Positive and 0= Negative			

Figure 2. shows the correlation coefficient and statistical distribution of the data among the independent features of ZADA dataset.

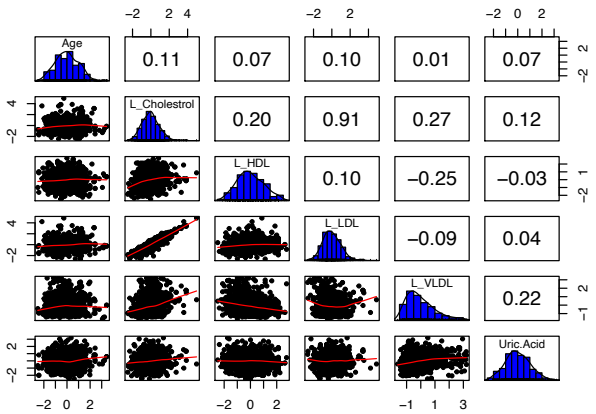

Figure 2. Correlation coefficient (upper triangle), histogram and density (diagonal), and statistical distribution of data (lower triangle) for the independent features of ZADA dataset

4.2 Classification Results

In our implementation for classification of ZADA dataset, our predictive evaluation uses the Bayesian point estimates from the training data. It takes the predictors from the testing data to calculate the predicted probabilities of the 0 (healthy) and 1 (diabetic) classes for each patient. Then, we compared the predicted to the actual outcomes to evaluate the classification model performance.

We used Bayesian logistic regression models with three different priors to build our Bayesian classifiers. To check the behaviour of the proposed models, we compared their performances with the traditional classifiers on the ZADA dataset. The model performances of the five classical algorithms along with the Bayesian models used are reported here. These models were used for predicting response "Diabetes" for both the training and testing datasets. Table 3 summarizes the values of evaluation metrics for each model, calculated from the confusion matrices.

Table 3. Performance of classification using classical and Bayesian algorithms for ZADA dataset

	Algorithms	ACC	Recall	Precision	F_1
Classical models	KNN	85.63	89.35	84.91	86.57
	DT	82.59	82.02	84.71	82.89
	NB	70.14	74.59	68.46	71.25
	LR	65.09	64.19	66.37	64.72
	SVM	61.98	50.08	65.01	56.68
Bayesian LR models	**Gaussian (0,10) prior**	**92.53**	**94.85**	**91.42**	**93.11**
	Laplace (0, 1) prior	91.72	89.92	90.95	91.43
	Cauchy (0,0.01) prior	90.81	89.45	89.66	89.56

Results in Table 3 show that the Bayesian models have given better model performance compared to the traditional classifiers. We can see that the highest accuracy, recall, precision and F1 of 92.53%, 94.85%, 91.42% and 93.11% respectively were obtained from BLR under Gaussian prior. Furthermore, amongst the Bayesian models investigated, the BLR model under Gaussian prior was the best one compared to the Laplace and Cauchy priors used for our dataset. We can also notice that the lowest accuracy measures were obtained from the SVMs algorithm, based on all evaluation metrics used, which indicates that the SVM model cannot be considered in classifying such kind of data. On the other hand, based on accuracy, recall, precision and F1 evaluation metrics, the best classical classifier was KNN with 85.63% accuracy, 89.35% recall, 84.91% precision, and 86.57% F1, respectively.

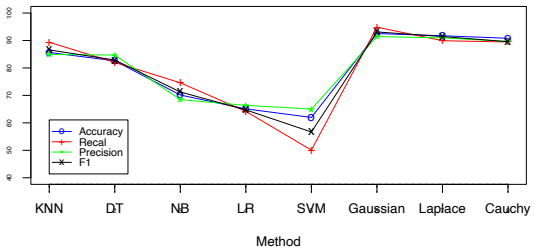

Figure 3. Performance of different classification algorithms using different evaluation metrics: Accuracy (**blue**), Recall (**red**), Precision (**green**), F1 score (**black**) for ZADA dataset.

As shown in Figure 3 and Table 3, all Bayesian models have performed better than the traditional models in classification performance. The model which performed best was the one based on Gaussian prior, providing accuracy= 92.53%, Recall=94.85%, Precision= 91.42%, and F1= 93.11% on the validation set. Amongst the three Bayesian models used, the one with Cauchy prior has the worst performance with an accuracy of 90.81%, recall of 89.45%, precision of 89.66%, and F1 of 89.56%. Although the results of the Cauchy model were not as good as the Gaussian model, it was better than almost all the other traditional models. This implies that the selected priors for our Bayesian modelling inference have a significant impact on the classification performances and have performed better than the classical approaches.

From the results obtained, we can see that all the Bayesian models investigated in this study have given better model performance compared to the traditional classifiers. This indicates that with proper prior distribution, the Bayesian models can make a significant improvement to the model performance, hence better predict diabetes. The main contribution was to investigate different priors and hence choose the most suitable one. We conclude that the Bayesian logistic regression with Gaussian prior had the highest performance for the ZADA datasets showing an accuracy of higher than 92%.

4.3 Bayesian Model Performance

Trace plots of samples are very important for checking the convergence assessment of the MCMC sampling algorithm, and the fluctuation of the trace plots show the equilibrium of the numerical distribution for the parameter. Figure 4 shows the posterior distribution for the coefficients of the model parameters under the Gaussian prior distributions. We can see that the densities of the posterior distributions are converged under the prior distribution used. The MCMC trace plots (left panels) are illustrating how well the samples are mixing, and the posterior densities (right panels) are providing a good understanding of the estimated coefficients from the Bayesian model. This indicates that the Bayesian inference is successful and the prior distributions used are accurate.

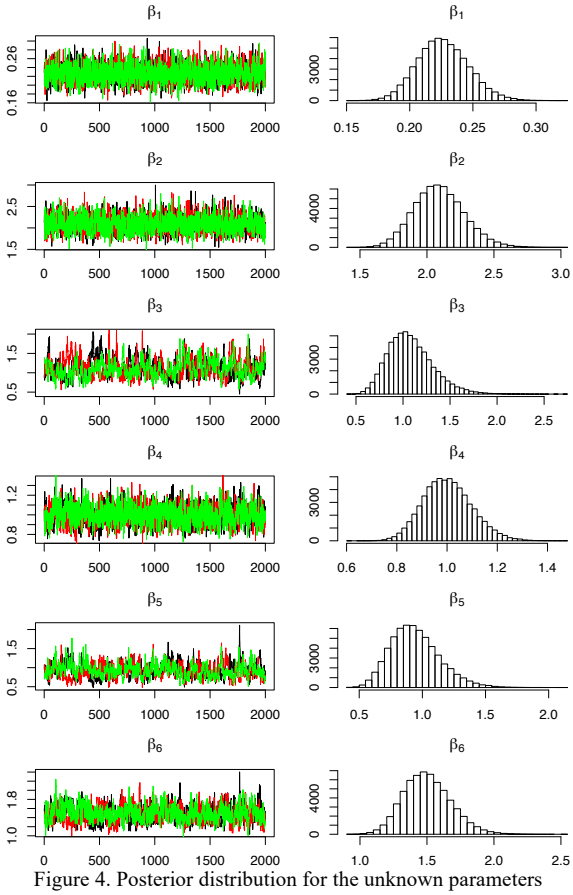

Figure 4. Posterior distribution for the unknown parameters ($\beta_1, \beta_2, \beta_3, \beta_4, \beta_5,$ and β_6) of the Bayesian logistic regression model under the Gaussian prior distributions. Left panels give the MCMC trace plots, and the right panels are posterior densities (histogram) for each parameter.

As we have seen, the choice of the best prior distribution is a challenging task in any Bayesian inference modelling. In this paper, we investigated the use of three different informative priors on the parameters of the logistic regression model. Results of the BLR model tend to provide good parameter estimation results under the priors used, which affected positively on the classification task. From the experimental results, we can see that although the Cauchy and Laplace prior distributions with location parameter 0 and different scale parameters have provided very good results, the Gaussian priors were better in the classification process. Therefore, using three different prior distributions have provided a very good understanding of the performance of the model when the prior distribution was changed. From using Gaussian prior distribution, the resulting posterior distribution for each parameter, given in Figure 4, follows an approximately normal distribution.

Table 4 shows the extracted corresponding posterior median estimates for the model coefficients. The uncertainties in our estimates are also calculated as the credible Bayesian intervals. Thus, the uncertainty intervals are calculated by finding the relevant quantiles of the samples from the posterior distributions, as given in columns 3 and 4 of Table 4.

Table 4. Posterior estimates of the model parameters using Gaussian prior

Attribute Name	Estimate Posterior (Median)	5%	95%
Intercept	0.2257	0.1945	0.2611
Age	2.0850	1.7973	2.4169
Cholesterol	1.0544	0.7366	1.5138
L_HDL	0.9980	0.8512	1.1709
L_LDL	0.9248	0.6577	1.2940
L_VLDL	1.4916	1.2348	1.8085
Uric Acid	0.6901	0.5964	0.7943

The Bayesian statistics in Table 4 obtained from posterior distributions have a simple way of managing the true values of model parameters in such a way that when the estimates of parameters contain zero in the calculated credible intervals, then it can be obtained from the final model.

It is worth noticing that the proposed Bayesian method has also been applied to the Pima dataset (Ravussin et al., 1994) as it is somehow similar to our ZADA dataset. The results of the classification of the Pima dataset were very similar to those of ZADA; thus, we have not reported the results of Pima experiments here. Therefore, our fully Bayesian approach leads to promising results based on statistically significant variables resulting from the posterior distributions when been implemented in different scenarios with different datasets.

5. DISCUSSION AND CONCLUSION

Logistic regression models have been widely used for classification problems in machine learning and data mining. In this paper, we presented a Bayesian approach for learning posterior distributions of logistic regression models using Gaussian, Laplace and Cauchy priors. The Bayesian models offer more flexibility and can handle more complex models, as they incorporate prior information into the analysis. With experimental studies, we have demonstrated that our MCMC algorithm can efficiently identify the uncertainty of the posteriors, and therefore obtained superior predictive performance. We have presented a fully BLR model that uses MCMC method for the classification of ZADA diabetes. All the uncertainties associated with the data have been incorporated into the analysis through the combination of prior knowledge and the likelihood of data via the Bayes theorem. Thus, the Bayesian model allows obtaining the statistical distribution of model parameters, instead of point estimations.

Experiments on diabetes classification problems demonstrated the performances of the proposed model. From the results of our experiments, we conclude that the Bayesian logistic regression with Gaussian prior had the highest performance for the ZADA datasets showing an accuracy of higher than 92%. It is interesting to observe from Table 4 that BLR with Gaussian prior has the highest evaluation performance. In all experiments, the traditional classifiers were not able to accurately predict all the cases. At the same time, the Bayesian models were able to classify almost all the patients as shown in Tables 4. The overall performance of the Bayesian logistic model according to all the evaluation measures has outperformed the traditional methods.

With these promising results of Bayesian models, the next step ought to study and examine the use of other priors for the proposed model, such as Student-t, Gamma, and Hyper-Lasso. In this study, we limited ourselves by only using three different priors. It would be more interesting to learn the impact of the choice of other (informative and non-informative) priors to investigate the classification of the ZADA dataset further. We would also like to observe the Bayesian results with non-informative priors along with some rare and uncommon priors to see how the model can be fitted. Another implementation of this study is that we only used two datasets with binary class outcomes here. We plan to extend our analysis to be used for multi-classes datasets on different application domains. Our future analysis will also evaluate the performance of Bayesian classification when using other priors such as Student-t and Hyper-Lasso.

REFERENCES

Chang, M., Dalpatadu, R. J., Phanord, D., Singh, A. K., Harrah, W. F., & Administration, H. (2018). *Breast Cancer Prediction Using Bayesian Logistic Regression*. *2*, 2–6. https://doi.org/10.31031/OABB.2018.02.000537

Clark, T. G., De Iorio, M., & Griffiths, R. C. (2007). Bayesian logistic regression using a perfect phylogeny. *Biostatistics*, *8*(1), 32–52. https://doi.org/10.1093/biostatistics/kxj030

DuMouchel, W. (2012). Multivariate bayesian logistic regression for analysis of clinical study safety issues. *Statistical Science*, *27*(3), 319–339. https://doi.org/10.1214/11-STS381

Gelman, A., Jakulin, A., Pittau, M. G., & Su, Y. S. (2008). A weakly informative default prior distribution for logistic and other regression models. *Annals of Applied Statistics*. https://doi.org/10.1214/08-AOAS191

Ghosh, J., Li, Y., & Mitra, R. (2018a). On the Use of Cauchy Prior Distributions. *Bayesian Analysis*, *13*(2), 359–383. https://doi.org/10.1176/appi.ajp.2014.13121571

Ghosh, J., Li, Y., & Mitra, R. (2018b). On the use of Cauchy prior distributions for Bayesian logistic regression. *Bayesian Analysis*. https://doi.org/10.1214/17-BA1051

Hassan, M. M., Jones, E., & Buck, C. E. (2019). A simple Bayesian approach to tree-ring dating. *Archaeometry*, *61*(4), 991–1010. https://doi.org/10.1111/arcm.12466

Hassan, Masoud M, & Amiri, N. N. (2019). Classification of Imbalanced Data of Diabetes Disease Using Machine Learning Algorithms. In G. E. Bostanci (Ed.), *ICTACSE 2019, 4th international conference of theoretical and applied computer science and engineering* (pp. 50–55). Istanbul, Turkey

Hassan, Masoud Muhammed. (2018). Bayesian Sensitivity Analysis to Quantifying Uncertainty in a Dendroclimatology Model. *ICOASE 2018 - International Conference on Advanced Science and Engineering*, 363–368. https://doi.org/10.1109/ICOASE.2018.8548877

Holmes, C. C., & Held, L. (2006). Bayesian auxiliary variable models for binary and multinomial regression. *Bayesian Analysis*, *1*(1 A), 145–168. https://doi.org/10.1214/06-BA105

Huggins, J. H., Campbell, T., & Broderick, T. (2016). Coresets for scalable Bayesian logistic regression. *Advances in Neural Information Processing Systems*, *Nips*, 4087–4095.

Joseph, L. (2016). *Bayesian Inference for Logistic Regression Parameters*. 1–12. https://doi.org/10.1111/j.1469-7793.2001.00521.x

Kass, R. E., Gilks, W. R., Richardson, S., & Spiegelhalter, D. J. (1997). Markov Chain Monte Carlo in Practice. *Journal of the American Statistical Association*. https://doi.org/10.2307/2965438

Li, L., & Yao, W. (2018). Fully Bayesian logistic regression with hyper-LASSO priors for high-dimensional feature selection. *Journal of Statistical Computation and Simulation*, *88*(14), 2827–2851. https://doi.org/10.1080/00949655.2018.1490418

Lin, J., Myers, M. F., Koehly, L. M., & Marcum, C. S. (2019). A Bayesian hierarchical logistic regression model of multiple informant family health histories. *BMC Medical Research Methodology*, *19*(1), 1–10. https://doi.org/10.1186/s12874-019-0700-5

Madigan, D., Genkin, A., Lewis, D. D., & Fradkin, D. (2005). Bayesian multinomial logistic regression for author identification. *AIP Conference Proceedings*, *803*(1), 509–516. https://doi.org/10.1063/1.2149832

Mary Gladence, L., Karthi, M., & Maria Anu, V. (2015). A statistical comparison of logistic regression and different bayes classification methods for machine learning. *ARPN Journal of*

Engineering and Applied Sciences, *10*(14), 5947–5953.

Maxime Vono, P. C. (2018). Sparse Bayesian Binary Logistic Regression Using The Split-And-Augmented Gibbs Sampler. *2018 IEEE International Workshop on Machine Learning for Signal Processing, Sept. 17–20, 2018, Aalborg, Denmark*.

Octaviani, T. L., Rustam, Z., & Siswantining, T. (2019). Ovarian Cancer Classification using Bayesian Logistic Regression. *IOP Conference Series: Materials Science and Engineering*, *546*(5). https://doi.org/10.1088/1757-899X/546/5/052049

Plummer, M. (2016). rjags: Bayesian graphical models using MCMC. In *R package version 3-13*. https://doi.org/http://cran.r-project.org/package=rjags

R Development Core Team, R. (2011). R: A Language and Environment for Statistical Computing. In *R Foundation for Statistical Computing*. https://doi.org/10.1007/978-3-540-74686-7

Ravussin, E., Valencia, M. E., Esparza, J., Bennett, P. H., & Schulz, L. O. (1994). Effects of a traditional lifestyle on obesity in Pima Indians. *Diabetes Care*. https://doi.org/10.2337/diacare.17.9.1067

Rstudio Team. (2019). RStudio: Integrated development for R. RStudio, Inc., Boston MA. In *RStudio*. https://doi.org/10.1007/978-3-642-20966-6

Spyroglou, I. I., Spöck, G., Chatzimichail, E. A., Rigas, A. G., & Paraskakis, E. N. (2018). A bayesian logistic regression approach in asthma persistence prediction. *Epidemiology Biostatistics and Public Health*, *15*(1), e12777-1-e12777-14. https://doi.org/10.2427/12777

Suleiman, M., Demirhan, H., Boyd, L., Girosi, F., & Aksakalli, V. (2019). Bayesian logistic regression approaches to predict incorrect DRG assignment. *Health Care Management Science*, *22*(2), 364–375. https://doi.org/10.1007/s10729-018-9444-8

Wang, H., Xiao, X., Zhang, X., Zhang, J., & Yan, Y. (2010). A bayesian logistic regression approach to spoken language identification. *IEICE Electronics Express*, *7*(6), 390–396. https://doi.org/10.1587/elex.7.390