Towards a Confidence-Centric Classification Based on Gaussian Models and Bayesian Principles

Dongxu Han, Hongbo Du and Sabah Jassim

Department of Applied Computing, University of Buckingham, Buckingham MK18 1EG, UK
{dongxu.han, hongbo.du, sabah.jassim}@buckingham.ac.uk

Abstract. Classification is an essential form of machine learning. It has been widely used in various application areas. Conventional classification schemes are mostly interested in the class label outcome rather than the strength of the class predictions. This paper introduced a new interpretation of "classification confidence" to complement the predicted class label. The level of confidence is formulated based on Gaussian probability models and the Bayesian classification principles. The paper shows the soundness of the concept and argues the very essence of embedding it in the centre of a classification process. The paper applies the proposed concept over a data set of miscarriage cases in early pregnancy and demonstrates that the concept works well and indeed reflects the degree of certainty regarding the data set.

Keywords: confidence, classification, Gaussian probability models, Bayesian principles

1 Introduction

Classification is one of the fundamental processes in machine learning. It has been widely applied in various applications. In a typical classification process, a model is trained by applying a learning algorithm to a set of training examples. The trained model is then applied to an unseen data sample to determine its class label. Often, a trained model is also tested on a set of testing examples to determine the level of accuracy of the model [1]. Various types of classification models and learning algorithms have been developed over the past several decades [2].

In recent years, classification techniques are increasingly deployed in medical diagnosis with promising results. Classification models are trained on various data features extracted from medical images and observations from medical tests. The models are then used in clinical examinations as a complementary tool for more accurate and timely diagnoses of diseases. In this type of real-life application scenarios, predicting the correct class label alone may not be sufficient since the classification decision must have adequate support especially when there are significant risky implications of misclassification. For example, in a scenario of diagnosing the right type of a tumour, it is important for doctors to know whether the tumour is benign or malignant *and* how much belief they have in that diagnostic decision. Such a requirement may also apply to many other application domains beyond medical diagnosis. Unfortunately, conventional approaches in classification are more interested in the predicted class label rather than how reliable the prediction is. The tested level of accuracy for a model can only provide a rough idea of the model's general reliability, but does not give indication about the level of certainty in each specific classification decision.

Various forms of similar concepts about classification confidence have been considered in the past. A k-nearest neighbour model may use the proportion of the majority in a majority voting scheme to indicate the level of confidence [3]. Decision tree classifiers can use the proportion of the majority of the training examples at a leaf node to determine the level of certainty [4]. But the preciseness of this type of measurement for confidence evaluation could be very much limited. A Support-Vector-Machine classifier may quantify its confidence by calculating the distance from the target data point to the decision hyper-plane [5], but this logic may have a limited range of application. In a recent work involving the second and the third authors [6], a simple criterion of confidence was developed. The scheme categorises each decision with a confidence band of either high, medium or low, based on thresholds that are set according to the boundaries between the classes, but these discrete bands are also limited. A similar but more formal approach to address this issue comes from conformal predictions [7]. In this approach, an error probability ε is introduced when a classifier is making decisions, and then the level of confidence can be seen as $1-\varepsilon$, i.e. the probability of correct classifications. It provides a very nice regression over discrete data samples, but modelling of the error rate ε by most likely following a Bernoulli distribution still seems quite naïve [8] [9].

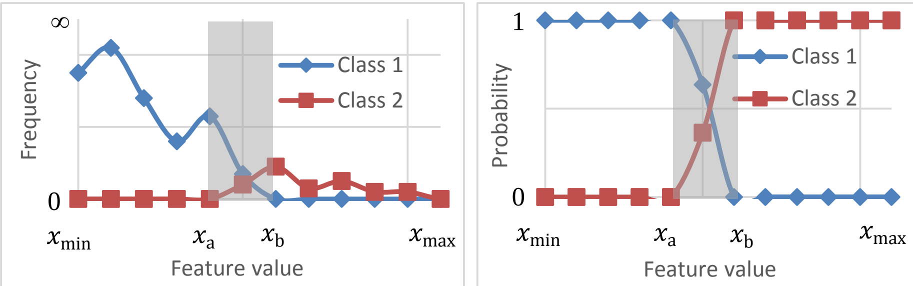
In this paper, we propose a new approach to address the issue of classification confidence from a slightly different angle of view than that in conformal predictions. The proposed approach defines the level of confidence as the difference between the posterior probability of predicting the class and the posterior probability of predicting other classes. The rationale behind this is quite straightforward: when the difference is large, the posterior probability of predicting one class is high, and hence the level of confidence is high for that class. When the difference is small, the posterior probabilities of predicting the possible classes are near to each other and hence the level of confidence about one of the classes should be low. In an extreme situation of equal probabilities, the level of confidence should be zero. The proposed approach is based on Gaussian models, and the definition of classification confidence is based on the Bayesian theorem. We shall present the rationale behind the concept and test its application by conducting an empirical study over a set of early pregnancy data. The evidence shows the validity of the concept as well as its practical potentials. We believe that the concept can be adopted not only with the outcome from one classifier but also in a confidence-based fusion framework in a complex decision support system.

The rest of the paper is organised as follows. Section 2 presents our definition of the classification confidence and the classification modelling in light of the confidence. Section 3 shows the experimental work of using the concept for quantifying prediction strength of miscarriage cases in early pregnancies. Section 4 discusses some related issues regarding the level of confidence. Section 5 concludes the paper with a summary of the work reported and future work needed.

2 The Proposed Method

2.1 Classification Confidence

In a typical training data set, examples of the individual classes may be distributed differently in the corresponding dimensional space. Figure 1 presents a simplified view of distributions of a set of one-dimensional training examples of two classes, and provides a conspicuous view about the strength of classification for each class. As illustrated by the frequency diagram in Figure 1(a), the two classes are very much distinct from each other when the data feature x has a value that is below a certain threshold x_a or above another threshold x_b due to the lack of examples from the opponent classes. However, conflicts of classification occur in a region between the two thresholds, where samples' feature of two classes starts to overlap. At the intersection point of the two curves, the overlapping occurs the most. Therefore, the overlapped region should be considered as the "zone of confusion" and the level of uncertainty in classifying sample attains the maximum value when the presences of the two classes are nearly equal.



(a) Frequency Distribution of Feature Values (b) Probability of a Class over Feature Values

Fig. 1. An Illustration of Value Distributions of Examples of 2 Classes

Based on this understanding, it is logical to transfer the previous frequency-based reasoning into a probability-based concept as shown in Figure 1(b), where the likelihood of the presence of different classes is a good indication of the confusion caused in classification. As illustrated in the diagram, all the discussed characteristics regarding "confusions" are well preserved with a normalised scale. The range between the two probability curves on the y-axis indicates the magnitude of the overlapping between the two classes, which should be considered as being proportional to the level of decision confidence.

Therefore, for a given finite class set $\{\omega\} = \{\omega_1, \omega_2, \dots, \omega_k\}$, the level of confidence can be presented as:

$$\mathrm{Confidence}(\omega_i|\vec{x}) \propto |\mathrm{P}(\omega_i|\vec{x}) - (1 - \mathrm{P}(\omega_i|\vec{x}))| \quad (1)$$

where $\mathrm{P}(\omega_i|\vec{x})$ is the conditional probability of predicting class ω_i based on a given feature vector $\vec{x}$ and therefore the aggregate probability of predicting into the rest of the classes will be $1 - \mathrm{P}(\omega_i|\vec{x})$.

The second term in the absolute difference in (1) is indeed the classification error rate ε for class ω_i at the given data point, which can be simplified as:

$$\text{Confidence}(\omega_i|\vec{x}) = |2\text{P}(\omega_i|\vec{x}) - 1| \quad (2)$$

This definition is justified by an assumption that the level of the confidence of the classification is directly proportional to the difference of the two probabilities without any transition bias, i.e. the gradient is equal to 1. Expression (2) motivates the introduction of a generalised confidence-centric score function for the classified label ω_i as:

$$\text{Decision score}(\omega_i|\vec{x}) = 2\text{P}(\omega_i|\vec{x}) - 1 \quad (3)$$

Here the sign of the decision score indicates the belonging of the class, which positive value would indicate a confirmation of the chosen class ω_i and negative value indicates a preference of the other classes. The absolute value of the decision score is the level of confidence in the decision made on the class belongings.

2.2 Decision Score Modelling

Here we introduce a Bayesian-based model for the decision score defined in equation (3). According to the Bayesian theorem:

$$\text{P}(\omega_i|\vec{x}) = \frac{\text{P}(\vec{x}|\omega_i)\text{P}(\omega_i)}{\text{P}(\vec{x})} \quad (4)$$

where $\text{P}(\omega_i)$ and $\text{P}(\vec{x})$ are two priors that represent the natural incidence of the class ω_i and the expected observation probability of feature $\vec{x}$, while $\text{P}(\vec{x}|\omega_i)$ is known as a posterior of the feature $\vec{x}$ given that it belongs to the class ω_i. Unfortunately, it is impossible to know exactly the priors in real-life scenarios due to unavoidable uncertainty and randomness. We, therefore, estimate the parameters by using the training dataset.

Given a sample space $\Omega = \{[\omega_1],[\omega_2],\ldots,[\omega_k]\}$, where $[\omega_i]$ is the set of all samples that belong to class ω_i, then $\text{P}(\omega_i)$ can be estimated as the proportion of the interested class ω_i to the total number of samples i.e.:

$$\text{P}(\omega_i) = \frac{|[\omega_i]|}{|\Omega|} \quad (5)$$

$\text{P}(\vec{x})$ and $\text{P}(\vec{x}|\omega_i)$ are the two probability functions describing the distribution of the feature $\vec{x}$, respectively within the overall population and within the population of class ω_i. Our proposed scheme assumes that both are Gaussian distributions. Consequently, a simplified model that is based on a single Gaussian distribution is proposed first as follows. Given the mean μ and variance σ^2 for a univariate feature $\vec{x}$, we use the Gaussian probability density function:

$$\mathcal{N}(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\sigma^2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

for characterizing $\text{P}(\vec{x})$ and $\text{P}(\vec{x}|\omega_i)$ as:

$$\begin{cases} P(\vec{x}) = \mathcal{N}(x \mid \mu_{\Omega}, {\sigma_{\Omega}}^2) \\ P(\vec{x}|\omega_i) = \mathcal{N}(x \mid \mu_{\omega_i}, {\sigma_{\omega_i}}^2) \end{cases} \tag{6}$$

In many applications, data features normally exist in a multidimensional space. Therefore, it is essential that we expand the previous simple model into a multivariate Gaussian model to accommodate multidimensional feature vectors. So for a given *d* dimensional data set with the mean vector $\vec{\mu}$ and covariance matrix Σ, we simplify the standard Gaussian probability density function $\mathcal{N}(\vec{x} \mid \vec{\mu}, \Sigma)$ as:

$$\mathcal{N}(\vec{x} \mid \vec{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi^d |\Sigma|}} e^{-\frac{(\vec{x} - \vec{\mu})\, \Sigma^{-1}\, (\vec{x} - \vec{\mu})^T}{2}}$$

We then derive $\vec{\mu}_x, \Sigma_x$ from $\omega = \{\omega_1, \omega_2, \dots, \omega_k\}$ and $\vec{\mu}_{\omega_i}, \Sigma_{\omega_i}$ from ω_i, and then $P(x)$ and $P(x|\omega_i)$ can then be characterized as:

$$\begin{cases} P(\vec{x}) = \mathcal{N}(\vec{x} \mid \vec{\mu}_{\Omega}, \Sigma_{\Omega}) \\ P(\vec{x}|\omega_i) = \mathcal{N}(\vec{x} \mid \vec{\mu}_{\omega_i}, \Sigma_{\omega_i}) \end{cases} \tag{7}$$

One concern is that we have tried only to use a single Gaussian in the modelling so far, but this may not always be realistic. Real-life data may reflect a combination of multiple Gaussians, each of which has its own mean vector and covariance matrix. Therefore, we have chosen to further extend our model into a Gaussian Mixture Model (GMM). In the mixture model, each sub-Gaussian model has been given a parameter set $\theta = \{W, \vec{\mu}, \Sigma\}$, where *W* represents the weight of each sub-model in the mixture and the summation of the weight of all the models should be 1. Therefore, given a sequence of *K* parameter sets $\{\theta_{i=1\dots K}\}$, i.e., *K* Gaussian sub models, we can characterize our universal mixture model as:

$$\mathcal{N}(\vec{x} \mid \theta_{i=1\dots K}) = \sum_{i=1}^{K} W_i \mathcal{N}(\vec{x} \mid \vec{\mu}_i, \Sigma_i)$$

Therefore, we would be able to derive parameter set θ_{ω_i} for each class from the relevant class set $\{\omega_i\}$ and the weight of each set would be considered as its proportion in the whole training set, i.e., $W_i = \frac{|[\omega_i]|}{|\Omega|}$, which $P(\vec{x})$ and $P(\omega_1)P(\vec{x}|\omega_i)$ can then be characterised as:

$$\begin{cases} P(\vec{x}) = \mathcal{N}(\vec{x} \mid \theta_{\omega_{i=1\dots k}}) \\ P(\omega_i)P(\vec{x}|\omega_i) = \mathcal{N}(\vec{x} \mid \theta_{\omega_i}) \end{cases} \tag{8}$$

3 Experiment Results

To evaluate the usefulness of the confidence concept introduced in the previous section, this study conducted several experiments using a collection of data about early pregnancies obtained from the Early Pregnancy Department, Queen Charlotte and

Chelsea Hospital, Imperial College London. The data collection contains three measurements of gestational sac sizes, i.e. major and minor diameters from the sagittal plane and major diameter from the transverse plane, recorded from the ultrasound machine. The collection also includes a variable, known as the Mean Sac Diameter (MSD), derived from the three diameter measurements. The collection consists of a training set of 94 examples (15 cases of miscarriage (MC) and 79 cases of PUV (Pregnancy of Unknown Viability)), and a test set of 90 examples (11 cases of MC and 79 cases of PUV)). Basically, PUV is declared when there are no clear signs of miscarriage although this may occur in subsequent scans.

We first trained the proposed decision score model and derived their parameters based on the training dataset. We then applied the proposed models on each testing example from the test set and measured the decision score for each testing example. As introduced in the proposed method section, the decision score would take a range of [-1, 1], which can be seen as the decision confidence towards [PUV, MC] in this binary class dataset.

3.1 Evaluation of the "Zone of Confidence"

Figure 2 presents scatterplots of the decision scores against feature values along the MSD dimension. According to the known literature in the related field of medicine, 25mm in MSD is a well-recognised threshold for separating PUV from miscarriage cases [10]. Therefore, we have rescaled the MSD dimension by setting (25, 0) as the origin, then plotted the related decision score of each feature value in the test set for each model accordingly. The corresponding class provided in the test set was plotted in markers of triangle and cross respectively. After the rescaling, the 1st, 2nd, 3rd and 4th quadrants in each scatterplot would indicate the possible classification results, i.e. true positive, false positive, true negative and false negative respectively. The figure shows that not only the confidence scores are very high for MC (≥31mm) and PUV (≤16mm), but also the existence of a confusion zone between 16mm and 31mm with the maximum confusion near the threshold of 25mm. This finding itself is interesting due to a well-known fact that 16mm was a previous threshold in use that is only recently being revived to 25mm because of concerns of potential false positives from some doctors [10]. This finding indicates that the confidence score does reflect the level of confidence in the diagnosis.

In addition, all the modelled data points were distributed according to a sigmoid pattern, which matches our expectation that the confidence would drop dramatically when it approaches the confusion point, i.e. the origin in the presented diagrams, or otherwise be stable at -1 or 1 when the feature value is outside the "confusion zone". The scatterplot also shows that the use of GMM has resulted in data confusions being moved from the false negative region into the false positive region.

Figure 3 shows the scatter plots of decision scores and feature values for multivariate situations. To clearly demonstrate the relationship between the decision scores and feature vector values, we on purpose combined the 3 diameter components of each feature vector into a single average value (in fact MSD), and display the location of the data point along the MSD dimension. At the same time, the decision scores are calculated using the original 3D feature vectors themselves.

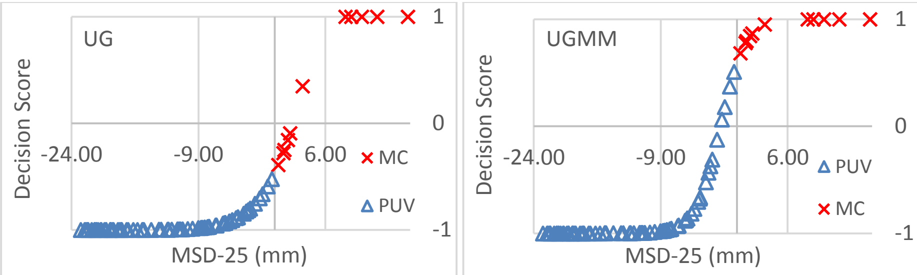


Fig. 2. Scatterplots of the MSD feature vs. decision scores in UG and UGMM situations (**UG: Univariate Single Gaussian on MSD; UGMM: Univariate Gaussian Mixture Model on MSD)*

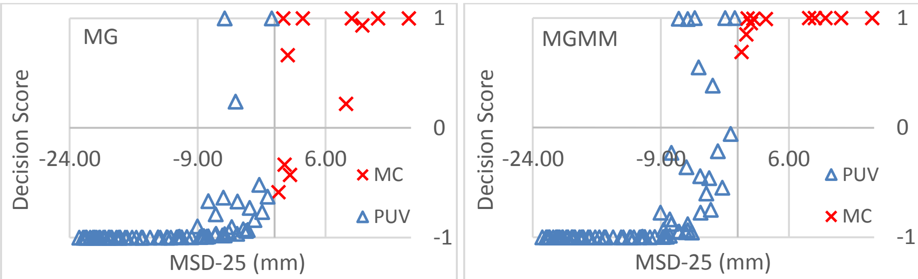


Fig. 3. Scatter plot of the feature values vs. decision scores in MG and MGMM situations (**MG: Multivariate Single Gaussian on 3 diameters; MGMM: Multivariate Gaussian Mixture Model on 3 diameters)*

Some general observations can be made from the scatterplots in both figure 2 and figure 3. The confusion zone clearly exists between the two thresholds, and the best fitting curve through the confusion zone tends to be close to a sigmoid line. The use of GMM also tends to move confusion cases from the false negative region into the false positive region, and at the same time increase the level of classification confidence in the true positive region. However, the confidence scores are more scattered in the confusion zone here than the scores for a univariate situation. The second scatterplot shows increased cases of misclassification even with high level of confidence.

In summary, the single Gaussian models tend to have a smoother fit to the sigmoid function, which reflected the nature of the proportional relationship between the MSD and the classification result. However, the performance of the multivariate Gaussian models shows that the data points are eventually made more distinguishable and pushed the classification results towards the two extremes, which provides a good sign for multivariate fusion in differentiating classes in highly overlapped feature values.

3.2 Evaluation of the Decision Score

As demonstrated in the previous section, the proposed method provides a good indication of the range of confidence/confusion. This study, however, also intends to evaluate the closeness of the decision score to a level of confidence exercised by manual diagnosis. The provided labels in the test set were the diagnosis results made by field experts based on known research truth.

Unfortunately, such a "manual diagnosis confidence" is not readily available in the data set provided. We, therefore, converted each of the given labels into to a decision score by mapping PUV and MC into -1 and 1 respectively, i.e. it is assumed that each of the decisions made by the experts was with absolute confidence. Following this, we would be able to evaluate the experiment result by calculating the difference between the human decision score and the derived decision score from the proposed confidence model. This difference would take a range of [0, 2], which 0 indicates a perfect match, and 2 would indicate an absolute conflict between the two decision scores. This evaluation was applied to each test example, and the average has been shown as an average error margin in Table 1.

Table 1. Confidence Error Margin Against Human Decision

Error Margin	*UG*	*UGMM*	*MG*	*MGMM*
MC	0.620	0.098	0.503	0.048
PUV	0.047	0.138	0.110	0.247
All	0.117	0.134	0.158	0.222

*UG: Univariate Single Gaussian on MSD; MG: Multivariate Single Gaussian on 3 diameters; UGMM: Univariate Gaussian Mixture Model on MSD; MGMM: Multivariate Gaussian Mixture Model on 3 diameters

The experiment result shows that SG best reflected the confidence nature of the PUV cases whereas MGMM had a better reflection on MC cases, which were only 0.047 and 0.048 units away from the human perspective on average. The experts had considerably more confidence on classifying MC cases base on well-understood diagnostic threshold [10]. Nevertheless, we have noticed that both SG and MG performed badly on fitting human perspectives on MC cases, which were 0.62 and 0.503 units less confident on average. It could be well caused by the domination of the PUV cases in the training data set, which makes the modelled natural expectation $\mathrm{P}(x)$ being very close to the probability of PUV classes.

This error can be well resolved by adopting GMM, which reduced the error margin of the SG and the MG to 0.098 and 0.048 respectively. However, the use of GMM may increase the error margin of the PUV cases at the same time. Nevertheless, it can still be tolerated since the influence was small, and after all, we should not expect the human experts being extremely confident on classifying PUV cases.

In general, the SG seems best fitting the human decisions. This is understandable since the real diagnoses made by the radiologists were based on the MSD measurement, which makes the model that use univariate features being more correlated to the evaluation result.

It is worth mentioning that the preciseness of the multivariate single Gaussian model and univariate GMM were fairly close to the best result (less than 0.1 units). Therefore, we can confirm that the multivariate based concept and the use of GMM are indeed valid and being effective for the fusion of multi-dimensional features, which we believe can be useful for extracting additional hidden knowledge from high dimensional space and cooperate better with wide class varieties.

4 Discussion

4.1 Intermedia function between confidence margin and decision score

In section 2.1, we have used the decision score in (3) as a simplified evaluation of the strength in decision making. However, the relation between the decision score and the final decision strength could be more flexibly treated due to the external bias involved during the decision making in the real practice. For example, if confusion exists in deciding over a scan image, the doctor may favour PUV over MC because PUV is a safer option, which has a second chance to be retested. Therefore, it would be helpful to introduce an intermedia function $\mathcal{T}$ that further refines the decision score S_D and yields a more realistic final decision towards the intended bias.

The regression of function $\mathcal{T}$ could be modelled in any format, which can be learned from a validation dataset after the probability model of the S_D has been trained. Take a linear regression model as an example; the transportation function $\mathcal{T}$ can be presented as:

$$\mathcal{T}(S_D) = m \cdot S_D + c : [-1,1] \mapsto [-1,1] \tag{10}$$

where *m* and *c* are the two constants that represent the transformation rate (a multiplier that indicates the trust of the calculated decision score) and the external bias (which one of the class we prefer more) respectively.

Unfortunately, this concept could not be validated due to the shortage of data samples. However, it is a valid idea and could enhance the preciseness of the confidence calculation.

4.2 The use of adaptive GMM under individual sub-classes

Arguably, as a further extension to (8), each class could also be modelled more realistically with a Gaussian mixture since it would be natural to have multiple subclasses in one class. However, the precise number of the sub-classes and their distributions are unknown. As a solution, we could first use Expectation Maximisation algorithm [11] to model the right distribution of each sub-Gaussian model of any given *K* values, and then selecting the best model set from all the possible *K* values by measuring their information criterions. Accordingly, we could derive the appropriate K_{ω_i} for each class

ω_i with their relevant parameter set $\{\theta_{\omega_{i_{j=0\ldots K}}}\}$, where our previous model $P(x)$ and $P(x|\omega_i)$ can then be further extended to:

$$\begin{cases} W_i = \frac{|[\omega_i]|}{|\Omega|} \\ P(\vec{x}) = \sum_{i=1}^{k} W_i \mathcal{N}(\vec{x} \mid \theta_{\omega_{i_{j=1\ldots K_{\omega_i}}}}) \\ P(\vec{x}|\omega_i) = \mathcal{N}(\vec{x} \mid \theta_{\omega_{i_{j=1\ldots K_{\omega_i}}}}) \end{cases} \quad (10)$$

We have in fact tried this way of modelling with the data set we have. However, we found that the *K* for each class is always equal to 1, which by nature result in a model that has no difference to the model in (8). Nevertheless, we believed that it is still worth to further validate this argument with different data sets in the near future.

5 Conclusions

This paper presented a concept of classification confidence, which reflects a systematic treatment of decision strengths based on posterior probabilities. The decision score function not only indicates the class label with +/− sign but also gives an absolute measurement of the level of confidence and belief to the decision made. The basic definition of confidence score based on a single univariate Gaussian distribution was extended to multivariate Gaussian mixture models. We argued the validity of the confidence score and tested its use in a real-life early pregnancy data set. The test result shows a strong correlation between the MSD feature values and the confidence-based predictions of miscarriage cases vs. PUV cases, indicating the potential use of the confidence level, which can be further explored in a multiple features and multiple classifier fusion framework.

Our future work includes further testing the use of confidence score in other real-life data sets of various kinds, fine tuning the concept as suggested in the Discussion part, and developing confidence-based fusion framework inside a decision support system.

Acknowledgment

The authors wish to thank the Early Pregnancy Department, Queen Charlotte and Chelsea Hospital, Imperial College London in general and Prof Tom Bourne and Dr Jessica Farren in particular for providing the data set for this study.

References

1. C. E. Metz, "Basic Principles of ROC Analysis," in *Seminars in Nuclear Medicine*, 1978.
2. M. Majnik and Z. Bosni, "ROC Analysis of Classifiers in Machine Learning: A Survey," 2011.

3. L. Xue, X. Yang, W. Wu, M. Chen and X. He, "An algorithm based on KNN and improved SVM for licence plate recognition," *Journal of Sichuan University,* vol. 43, no. 5, pp. 1031-1036, 2006.
4. H. Du, Data Mining Techniques and Applications, An Introduction, Cengage Learning, 2010.
5. R. Li, S. Ye and Z. Shi, "SVM-KNN classifier-a new method of improving the accuracy of SVM classifier," *Acta Electronica Sinica,* vol. 30, no. 5, pp. 745-748, 2002.
6. S. Khazendar, H. Al-Assam and H. Du, "Automated Classification of Static Ultrasound Images of Ovarian Tumours Based on Decision Level Fusion," in *Computer Science and Electronic Engineering Conference (CEEC)*, 2014.
7. G. Shafer and V. Vovk, "A Tutorial on Conformal Prediction," *Journal of Machine Learning Research* , pp. 371-421, 2008.
8. J. Wang, P. Neskovic and L. N. , "A Statistical Confidence-Based Adaptive Nearest Neighbor Algorithm for Pattern Classification," in *International Conference on Machine Learning and Cybernetics*, 2006.
9. H. Papadopoulos, V. Vovk and A. Gammermam , "Conformal Prediction with Neural Networks," in *19th IEEE International Conference on Tools with Artificial Intelligence*, 2007.
10. T. Bourne and C. Bottomley, "When is a pregnancy nonviable and what criteria should be used to define miscarriage?," *Fertility and sterility,* pp. 1091-1096 , 2012.
11. T. K. Moon , "The expectation-maximization algorithm," *IEEE Signal Processing Magazine,* vol. 13, no. 6, pp. 47 - 60 , 1996.