IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received October 23, 2016, accepted November 17, 2016, date of publication January 9, 2017, date of current version January 23, 2017.

Digital Object Identifier 10.1109/ACCESS.2016.2647238

A Hybrid Feature Selection With Ensemble Classification for Imbalanced Healthcare Data: A Case Study for Brain Tumor Diagnosis

SHAMSUL HUDA[1], JOHN YEARWOOD[1], HERBERT F. JELINEK[2], (Member, IEEE), MOHAMMAD MEHEDI HASSAN[3], (Member, IEEE), GIANCARLO FORTINO[4], AND MICHAEL BUCKLAND[5]

[1]School of Information Technology, Deakin University, Burwood VIC 3128, Australia
[2]School of Community Health, Charles Sturt University, Sydney NSW 2127, Australia
[3]College of Computer and Information Sciences, King Saud University, Riyadh 11543, Saudi Arabia
[4]Department of Informatics, University of Calabria, 87036 Rende, Italy
[5]Discipline of Pathology, School of medical Sciences, The University of Sydney, Sydney NSW 2006, Australia

Corresponding author: M. M. Hassan (mmhassan@ksu.edu.sa)

This work was supported by the Deanship of Scientific Research at King Saud University through the Research Group under Project RGP-281.

ABSTRACT Electronic health records (EHRs) are providing increased access to healthcare data that can be made available for advanced data analysis. This can be used by the healthcare professionals to make a more informed decision providing improved quality of care. However, due to the inherent heterogeneous and imbalanced characteristics of medical data from EHRs, data analysis task faces a big challenge. In this paper, we address the challenges of imbalanced medical data about a brain tumor diagnosis problem. Morphometric analysis of histopathological images is rapidly emerging as a valuable diagnostic tool for neuropathology. Oligodendroglioma is one type of brain tumor that has a good response to treatment provided the tumor subtype is recognized accurately. The genetic variant, 1p-/19q-, has recently been found to have high chemosensitivity, and has morphological attributes that may lend it to automated image analysis and histological processing and diagnosis. This paper aims to achieve a fast, affordable, and objective diagnosis of this genetic variant of oligodendroglioma with a novel data mining approach combining a feature selection and ensemble-based classification. In this paper, 63 instances of brain tumor with oligodendroglioma are obtained due to prevalence and incidence of the tumor variant. In order to minimize the effect of an imbalanced healthcare data set, a global optimization-based hybrid wrapper-filter feature selection with ensemble classification is applied. The experiment results show that the proposed approach outperforms the standard techniques used in brain tumor classification problem to overcome the imbalanced characteristics of medical data.

INDEX TERMS Brain tumor, morphological features, ANNIGMA, MRMR, feature selection, classification.

I. INTRODUCTION

Due to the widespread use of Electronic Health Records (EHRs) in many healthcare facilities, healthcare data are available for analysis in order to improve the quality of patient care more efficiently. However, exploration of healthcare/medical data is challenging due to its inherent heterogeneity, incompleteness, unbalanced and high dimensional nature. Often medical data are heterogeneous where the patients' recordings have different types of values, including real and integer with different ranges, image and text types. Most of the time, collection of medical data is not done purposely, instead; the data come as a by-product from the health care system. Due to the many dangerous and cost-sensitive natures of the diagnostic tests, components of data and related diagnosis tests are not always completed unless it is strictly required. Often the classes of patients who have the disease are significantly less than who don't have the disease. Therefore, data can be incomplete and unbalanced inherently. Finally, the data analysis approach should be able to interpret the results of analysis; a black box technique is highly unlikely to be accepted by the practitioners in a healthcare system. To deal with this challenge, the demand

of advanced data driven and machine learning techniques is constantly increasing. This paper addresses the issue of imbalanced healthcare/medical data through a case study using real brain tumor diagnosis problem.

Morphometric analysis of histopathological images is rapidly emerging as a valuable diagnostic tool for a variety of diseases [1]–[5], especially true tumor automated diagnostics. Enhanced computer imaging and analysis are paving the way for programs such as PAPNET, which is a program designed to achieve automatic diagnosis of cervical cancer through morphometric analysis of pap smears [5].

Oligodendroglioma is a subset of brain tumors, which has a high rate of responsiveness to chemotherapy. A correlation between chemosensitivity and a genetic variation of oligodendrogliomas on two particular alleles has recently been observed [6], [7]. The total loss of chromosome arms 1p and 19q is defined as ''1p-/19q'' and is known as the genetic variation [6]. It increases the chemosensitivity of the tumor to treatment and can lead to a better outcome [7].

The current gold standard in diagnosing oligodendrogliomas is histopathological classification. This technique requires extensive subjective decision making based on histological characteristics seen in the prepared slides and has led to a wide variance in diagnosis of oligodendrogliomas and a subsequent high degree of uncertainty in the incidence and prevalence of the tumor. Molecular testing of 1p/19q co-deletion with fluorescent in situ hybridization provides an accurate diagnosis but is an expensive technique and in most countries without a highly-developed medical system not available.

The 1p/19q co-deletion has characteristic features on hematoxylin-Eosin (H&E) stained tumorigenic slides, which are indicative of an oligodendroglioma [8]. These include round homogeneous nuclei, chicken wire like vasculature and perinuclear halos (cytoplasmic clearing). The chromosome deletion can result in a different blood vessel distribution and a deviated boundary property of the nuclei by influencing the structural organization of the chromatin within the nucleus [10]. A recently conducted study by Scheie investigated the association between ten such histological variables, location and genetic losses at 1p, 19q [9]. They found that the most significant feature was the round homogeneous nuclei. A more recent pilot study confirmed the ability of automated feature analysis combined with data mining to identify the 1p,19q variant [52], [53].

Histological classification may also be performed by Weber local descriptor image analysis and Grey value co-occurrence statistics. However, they are not reliable enough for H&E stained oligodendroglioma classification [11]. Wavelet-based measures have also been described throughout the literature [12], [13].

Image texture as a function of the grey scale values that make up the image texture can also be analysed using fractal geometry [14]. Fractal analysis [15]–[18] has been used for classifying malignant and nonmalignant tumors. Results confirm the scale invariance of 2D grey scale histological images and the suitability of chaos theory to analyze them. In order to find the difference between the images collected from non-malignant and malignant mammary tumors, correlation dimension and Higuchi's dimension analysis have been used in [14] and [18]. However, a major drawback of the available data collection and diagnosis approach is that brain tumor data from patients are imbalanced and smaller in size.

This study aimed to achieve fast, affordable and objective diagnosis of an oligodendroglioma. More specifically the study aimed to identify oligodendrogliomas with the chemosensitive attribute of co-deletion on the 1p/19q alleles using a set of morphological features, including fractal analysis from an unbalanced small data and combining this with novel data mining algorithms.

This paper proposes a globally optimized hybrid siginificant tumor feature selection algorithm which is combined with bagging and decision tree to generate an interpretable set of simplified diagnosis rule for tumor classification. The hybrid tumor feature selection combines a wrapper heuristic score computed from a Globally optimized Artificial Neural Network Input Gain Measurment Approximation (GANNIGMA) score [47] which is derived from an Artificial Neural network (ANN) based wrapper with a filter heuristic score. Global optimization in the training of wrapper and thereby, computation of the hybrid feature score in the proposed feature selection and ensemble technique in classification enhances the performance of diagnostic classifier to overcome the shortcoming of imbalanced data set.

The organization of the paper and description of different sections are as follows. A review of the related literature is provided in Section 2. Proposed approach for hybrid feature selection and ensemble classification to aid brain tumor diagnosis is described in Section 3. Results are reported and discussed in Section 4. Conclusions are made in Section 5.

II. RELATED LITERATURE

Computationally intelligent techniques and data mining approaches have increasingly been used for disease diagnosis [19]–[26], [55]–[59]. Many approaches and techniques have been reported for automatic classification of brain tumors, mostly in Computed Tomography (CT) and Magnetic Resonance Imaging (MRI) [27]–[33]. Classification methods included support vector machine (SVM) [34], [35], neural network [36], knowledge-based techniques [37], expectation–maximization (EM) algorithms and Fuzzy c-means (FCM) clustering.

An approach for classification of brain tumor tissues into normal, benign or malignant tumors was discussed in [34]. These authors used magnetic resonance images (MRI) and applied SVM and genetic algorithms (GA). They used the most common kernel functions, including linear, polynomial of various degrees and Radial Basis Function (RBF) [38]. The input to the SVM algorithm where the feature subset selected using GA during the data pre-processing step. In this approach, the accuracy of identifying any type of

tumor varied from 94.44 to 98.14%. No detail was given of the dataset and the type of tumors included in the analysis.

Hum *et al.* [39] applied GA and data mining techniques for classification of glioma subclasses and extracted histological features from atomic force microscopy (AFM) images. Results were computed overall possible parameter configurations, and computational experiments performed 100 runs to overcome stochasticity of results. The algorithm was able to distinguished grade II tumors (low-grade gliomas, which grow slowly) from grade IV tumors with a classification accuracy of 94.74%. However, the oligodendroglioma brain tumor was not included in the analysis [39].

Similarly, Papageevgious and Spyridonos [41] proposed a brain tumor grading model using fuzzy cognitive maps but did not include ologodendrogliomas. Fuzzy cognitive maps used to represent and model experience, expertise and heuristic of experts obtained a diagnostic output accuracy of 90.26% & 93.22 % for low grade and high-grade brain tumor respectively.

Nabizadeh and Kubat [42] have developed a highly accurate, low computational cost and fully automated system based on statistical features and compared this to Gabor wavelet features using several classifiers to detect portions of a tumor and delineate the tumor area from MRI images. An artificial neural networks-based method to detect brain tumor tissue from MRI images. A comparative effectiveness of the performances of statistical features and Gabor wavelet features for different wrapper classifiers has been accomplished in their study [42]. They claimed that their technique performs effectively in segmenting brain tumor tissues, which provide high classification accuracy and has a required low computational cost. But their proposed work did not include grading the tumors to classify Oligodendroglia.

Kharat *et al.* [43] proposed an artificial neural network based method to find out abnormalities of brain tumor on MRI images. The ANN in [43] was a combination of feed-forward and feedback propagation neural networks, including a number of processing steps such as segmentation of an image, extraction of features from the images and learning of a model from the training image data. The research with the highest reported accuracy of 99% for detection of brain tumors was reported using MRI and naïve Bayes classification in [44]. The tumor region was extracted by applying boundary detection methods and K-means clustering. However, none of the above methods considered oligodendrogliomas. As the 1p,19q variant has a good treatment outcome compared to other gliomas but requires specific treatment, it is very essential to accurately detect this tumor variant and distinguishes it from other gliomas.

III. METHODOLOGY

A. HISTOLOGICAL CHARACTERISTICS AND BRAIN TUMOR DIAGNOSIS PROBLEM

The data collection procedure from different patients was approved by the University of Sydney Human Ethics Committee (HREC# 12353). Neurology patients at the Royal Prince Alfred Hospital, Sydney Australia with and without the 1p19q co-deletion variant of ologodendroglioma was identified by neuropathologists following an autopsy and genetic testing with fluorescent in-situ hybridisation (FISH) and following the recommended procedure of a manufacturer of commercial Vysis FISH BAC probes (1p36/1p25 + 19q13/19p13 FISH probe kit) (Abbott Molecular, USA). Pathology samples cut at 5 micron thickness were fixed in formalin and embedded in paraffin before staining with hematoxylin and Eosin to view nuclei and surrounding cell details within the tissue. The Zeiss AxioCam HR camera linked to a Zeiss Axioxope A.1 microscope was used to conduct the image acquistion. The the collected images were analysed in triplicate at 40X magnification at 300 dpi resolution on a Zeiss Axiovision 4.8 (Zeiss, Germany) for the required histological characteristics and 1p,19q co-deletions confirmed. Morphological features were obtained with standard image analysis software (Metamorph V7.6.4.0, Molecular devices, CA). From the pathology slides the nuclei were segmented. Then different tumor features have been extracted from the segmentation, including equivalent sphere surface area, shape factor, orientation, height, width, inner radius, mean radius, outer radius, equivalent radius, area, and equivalent sphere volume. A sample image has been presented in Fig 1.

B. PROPOSED GANNIGMA BASED HYBRID FEATURE SELECTION WITH ENSEMBLE TECHNIQUE

The proposed approach develops a globally optimized Artificial Neural Network Input Gain Measurment Approximation (GANNIGMA) based hybrid feature selection which is combined with an ensemble classification (GANNIGMA-ensemble) technique to generate the diagnostic decision rule. The GANNIGMA hybrid feature selection in the proposed approach finds the significant features which help to generate a simplified rule. Ensemble classifier improves the classification accuracy. The feature selection based ensemble frame work is presented in Fig 2.

1) GANNIGMA HYBRID FEATURE SELECTION USING GLOBAL OPTIMIZATOION AND ARTIFICIAL NEURAL NETWORK (ANN)

Filter approach can find the intrinsic relationships between the individual diagnosis feature and tumor class. However, filter approach did not use any performance evaluation criteria based on accuracies. The filter is computationally cheap but does not ensure that the selected final tumor feature set would be the most significant in terms of performance. In contrast, wrapper approach uses accuracy based performance evaluation. Since the wrapper approach uses a classification accuracy based performance evaluation criteria during training, it can be ensured from wrapper approach that selected subset by the wrapper can achieve a better peformance; however, it may take more computational cost. The proposed hybrid approach integrates the knowledge about the intrinsic relationship between a particular feature with corresponding class estimated by the filter in the wrapper

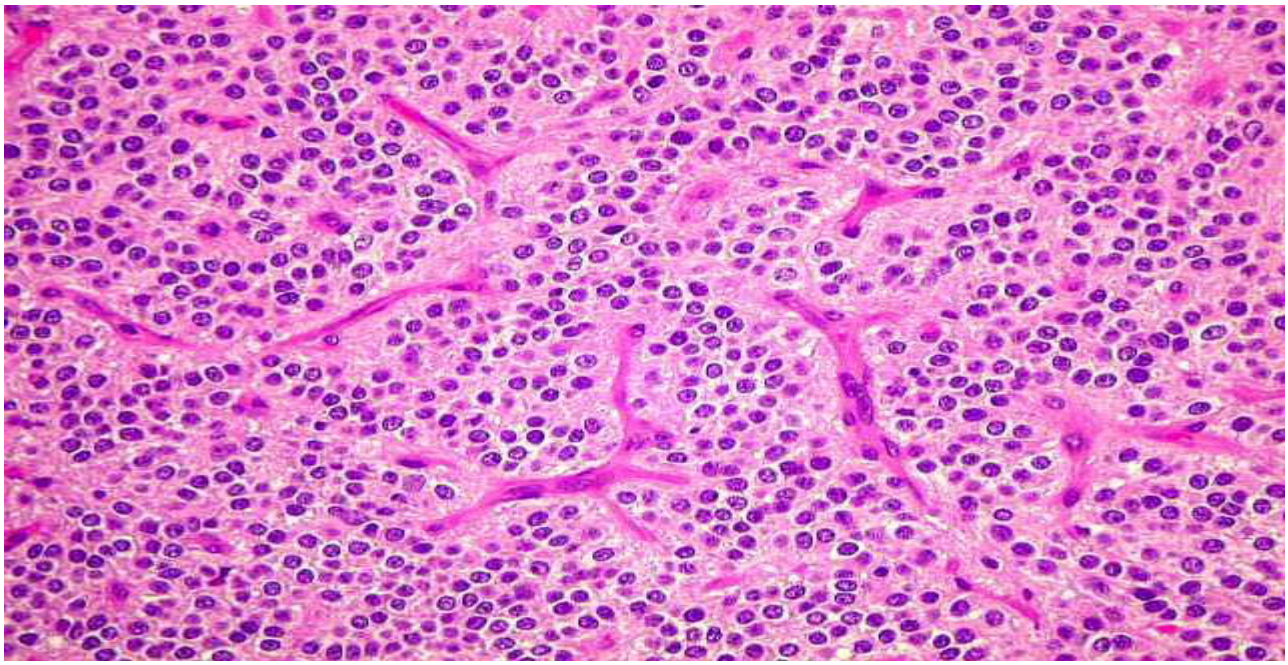

FIGURE 1. H&E stained histology slide showing perinuclear halo and chicken wire-like vasculature (40X magnification).

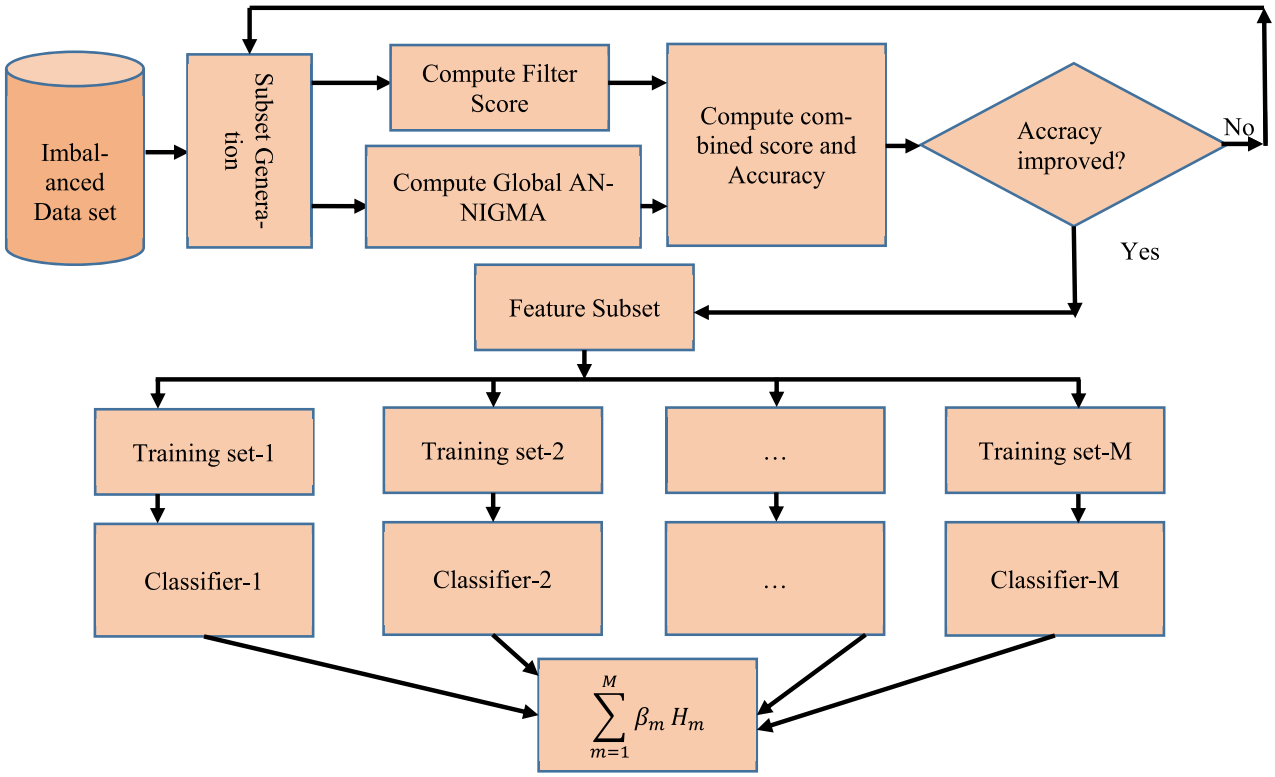

FIGURE 2. Proposed feature slection with the ensemble classification framework.

search process and takes advantages of the complementary properties of both approaches.

In our proposed tumor feature selection approach, a mutual information (MI) based Maximum Relevance Minimum Redundancy (MRMR) [47], [48] filter ranking heuristic is combined with the wrapper heuristic. The wrapper is taken as an Artificial Neural Network (ANN) [47], [48]. A wrapper heuristic is computed, which is Artificial Neural Network

Input Gain Measurement Approximation (ANNIGMA) [47], [48]. The hybridization of wrapper and filter is presented in the top part of Fig.2.

Improvements for both approaches were achieved by incorporating the filter feature ranking score with the wrapper approach to speed up the search process. Then an induction algorithm is used during the wrapper training process for selecting the optimal feature subset. The maximum relevance (MR) [48] algorithm contributed to redundancy while selecting the features that are highly relevant to class but highly correlated. Therefore, a redundancy function is incorporated (MR-*Minimum Redundancy*; *MRMR*) into the MR algorithm as in Eq. (1).

$$MRMR = \frac{1}{\max s|S|}\sum\nolimits_{f_p \in S} I\left(F_p; c\right) - \frac{1}{|S|^2}\sum\nolimits_{p,q \in S} I(F_p; F_q) \quad (1)$$

Where I(Fp;Fq) is the mutual information between the features Fp and Fq.

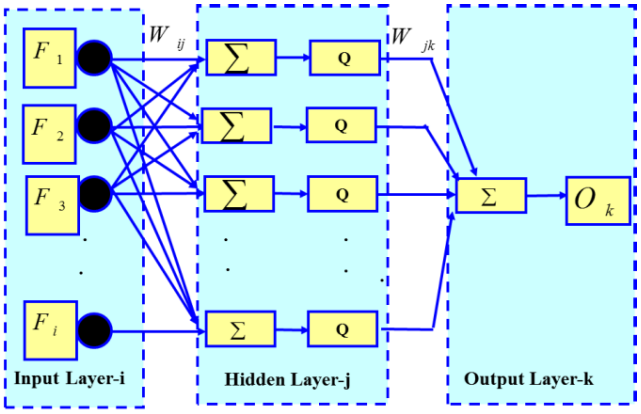

FIGURE 3. Multilaer Perceptron in Artificial Nneural Network.

2) COMPUTATION OF GLOBAL ANNIGMA (GANNIGMA) SCORE

Artificial Neural Network Input Gain Measurement Approximation (ANNIGMA) [47] is computed from the wrapper training of ANN. In general, a three-layer ANN has a structure as presented in Fig 3. Let us assume that the input, hidden and output layer are denoted as i, j, k. The logistic activation function is denoted as "Q" as below

$$Q(z) = (1/(1 + \exp(-z))) \quad (2)$$

A linear function is considered for first and second layer, then network output is as (3). Here F_i are the input tumor feature

$$O_k = \sum_j Q\left(\sum_i F_i \times W_{ij}\right) \times W_{jk} \quad (3)$$

The local gain for input tumor feature F_i is as (4)

$$LG_{ik} = \frac{\Delta O_k}{\Delta F_i} \quad (4)$$

The local gain is transformed as defined in [47] and (5):

$$LG_{ik} = \sum_j |W_{ij} \times W_{jk}| \quad (5)$$

For feature-i (F_i) is the local gain (LG) normalized based on a unity scale as (6) which is the ANNIGMA.

$$ANNIGMA(F_i) = \frac{LG_{ik}}{\max imum\ (i) LG_{ik}} \quad (6)$$

The standard back propogation training algorithm of ANN provides locally optimized parameters, which could be worse for imbalanced dataset. Therefore, a global optimization approach has been adopted with the standard backpropagation training. An Algorithm for Global Optimization Problem (AGOP) proposed in [45] and [46] is applied for optimal estimation of ANN parameters in the training of ANN. An average optimal wrapper heuristic for a Global ANNIGMA (GANNIGMA) scores is computed using an n-fold cross validation during the training of ANN.

The GANNIGMA score over cross-validation is computed as Eq. (7):

$$GANNIGMA(F_p)_{average} = \left(\frac{1}{n}\right)(GANNIGMA\ \left(F_p\right)_1 + \ldots + (GANNIGMA\ \left(F_p\right)_n \quad (7)$$

The maximum relevance score of a candidate tumor feature in a candidate subset and redundancy score beteween the candidate feature from the rest of the subset are used to calculate the Maximum Relevance and Minimum Redundancy (MRMR) score [48]–[50] which have been shown in Eq. (8):

$$\max_{F_i \in F - F_{l-1}} \left(\frac{1}{|S|}\sum_{f_i \in S} I(F_i; c) - \frac{1}{l-1}\sum_{F_j \in F_{l-1}} I(F_i; F_j)\right) \quad (8)$$

Since the combination of the candidate features in the subset and rest of the feature can be very large, an incremental search approach [49] has been used to calculate the MRMR score for candidate tumor feature. Then the features are ordered based on an equivalent weighted score which is presented in Eq. (9):

$$\text{Weighted MRMR}\left(\mathrm{F_p}\right) = 1 - \left(\frac{\text{Rank feature}\left(\mathrm{F_p}\right)\ \text{in MRMR}}{|\mathrm{F}|}\right) \quad (9)$$

A combined score for the MRMR-GANNIGMA hybrid approach is finally computed as in Eq. (10):

$$Combined\ Score\ (MRMR{-}GANNIGMA; F_P) = Weighted\ MRMR\ Score\ (F_P) + GANNIGMA(F_P)_{average} \quad (10)$$

In the wrapper stage a Multilayer Perceptron (MLP) Network (as illustrated in Fig. 3) [47], [48] is used, which is trained by combining AGOP and an n-fold cross-validation approach.

In the BE iteration, a number of trial runs has been executed with the n-fold cross validation. The average classification accuracy of the trial runs with its corresponding n-fold of the wrapper has been used to evaluate the feature subset. A backward elimination (BE) process updates MRMR and ANNIGMA [47], [48] and the combined score in every iteration. The tumor feature with the lowest combined score is excluded from the candidate set in each iteration, and the iterative process is continued for a cardinality of the candidate feature set equals to one. The subset with the highest accuracy and with the least cardinality is then chosen as the final feature subset.

3) ENSEMBLE CLASSIFICATION AND RULE GENERATION

Following the feature selection, classification of the test examples is performed using decision tree in combination with Bootstrap aggregating or bagging [50] machine learning algorithms. Bagging is a simple algorithm which uses bootstrap sampling. Given a training dataset T containing n examples, a sample of training examples, T_m, where m is 1 to M is created by selecting n examples uniformly at random with replacement from T(some examples can be selected repeatedly while some may not be selected at all). A particular classifier $H_m : m = 1.., M$ is learned based on the actual training set T_m. Then a compound classifier (H) is created by aggregating the particular classifiers. A new instance t_i is then classified to class c_j according to the number of votes obtained from particular classifiers H_m as in Eq. 11.

$$H\left(t_i, c_j\right) = sign\left(\sum_{m=1}^{M} \beta_m H_m\left(t_i, c_j\right)\right) \tag{11}$$

Where parameters $\beta_{m:}m = 1, \ldots, M$ are determined to optimize the final prediction by selecting the most accurate classifiers.

A decision tree is a popular data mining approach which focuses on creating a decision rule generation model. Decision tree can predict the value of a target feature by

TABLE 1. Classification accuracy obtained using GANNIGMA feature selection with MRMR, AGOP, bagging and decision tree.

TP Rate	FP Rate	Precision	F-Measure	ROC Area	Class
0.742	0.438	0.622	0.676	0.636	Y
0.563	0.258	0.692	0.621	0.636	N
0.651	0.346	0.658	0.648	0.636	Weighted Avg.

TABLE 2. Comparison of classification accuracy for different combinations of feature selection and classification techniques.

Techniques	TP Rate	FP Rate	Precision	F-Measure	ROC Area
All feature Decision Tree	0.444	0.548	0.434	0.414	0.471
All Feature SVM	0.381	0.624	0.367	0.367	0.379
ANNIGMA+ Decision Tree	0.429	0.577	0.417	0.412	0.392
ANNIGMA+ SVM	0.349	0.653	0.347	0.347	0.348
GANNIGMA + Decision Tree	0.492	0.508	0.492	0.492	0.507
ANNIGMA+ MRMR+ Decision Tree	0.508	0.484	0.517	0.472	0.542
GANNIGMA+ MRMR+ Decision Tree	0.556	0.437	0.579	0.527	0.588
All feature Bagging	0.524	0.479	0.523	0.519	0.559
GANNIGMA+ Bagging+ Decision Tree	0.571	0.427	0.573	0.571	0.521
ANNIGMA+ MRMR+ Bagging	0.54	0.462	0.539	0.538	0.568
ANNIGMA+ MRMR+ SVM+ Bagging	0.397	0.602	0.397	0.397	0.367
GANNIGMA+ MRMR + Bagging+ Decision Tree	0.651	0.346	0.658	0.648	0.636

```
InnerRadius <= 3.91754
|  Dh_x_20 <= 1.749892: Y
|  Dh_x_20 > 1.749892
|  |  Height <= 24.981308: N
|  |  Height > 24.981308: Y
InnerRadius > 3.91754
|  Width <= 26
|  |  OuterRadius <= 13.263317
|  |  |  OuterRadius <= 10.982136: N
|  |  |  OuterRadius > 10.982136
|  |  |  |  Dh_x_20 <= 1.699788
|  |  |  |  |  Dh_x_20 <= 1.663709: Y
|  |  |  |  |  Dh_x_20 > 1.663709: N
|  |  |  |  Dh_x_20 > 1.699788: Y
|  |  OuterRadius > 13.263317
|  |  |  OuterRadius <= 16.088008: N
|  |  |  OuterRadius > 16.088008: Y
|  Width > 26: Y
Tree size=19
```

(a)

```
ShapeFactor < 0.311657: Y
ShapeFactor >= 0.311657
|  OuterRadius < 11.101213: N
|  OuterRadius >= 11.101213
|  |  ShapeFactor < 0.5902
|  |  |  OuterRadius < 15.684648: N
|  |  |  OuterRadius >= 15.684648
|  |  |  |  OuterRadius < 16.4602015
|  |  |  |  |  Dh_x_20 < 1.8332585
|  |  |  |  |  |  Dh_x_20 < 1.740415
|  |  |  |  |  |  |  OuterRadius < 16.0017: Y
|  |  |  |  |  |  |  OuterRadius >= 16.0017: N
|  |  |  |  |  |  Dh_x_20 >= 1.740415: Y
|  |  |  |  |  Dh_x_20 >= 1.8332585: N
|  |  |  |  OuterRadius >= 16.4602015: N
|  |  ShapeFactor >= 0.590199: Y

Tree size=17
```

(b)

```
ShapeFactor <= 0.310617: Y
ShapeFactor > 0.310617
|  Height <= 23.580645
|  |  Area <= 329.907895
|  |  |  MeanRadius <= 9.450369
|  |  |  |  OuterRadius <= 10.982136: N
|  |  |  |  OuterRadius > 10.982136
|  |  |  |  |  Orientation <= -8.61876
|  |  |  |  |  |  Area <= 261.933333: Y
|  |  |  |  |  |  Area > 261.933333: N
|  |  |  |  |  Orientation > -8.61876: Y
|  |  |  MeanRadius > 9.450369: N
|  |  Area > 329.907895: Y
|  Height > 23.580645
|  |  OuterRadius <= 16.097414: N
|  |  OuterRadius > 16.097414: Y

Tree size=17
```

(c)

FIGURE 4. Deceision tree computed by proposed approaches with the ensemble technique. (a) GANNIGMA+ Bagging + decision Tree. (b). GANNIGMA + MRMR + Decision Tree. (c). All Feature+ Bagging + decision Tree.

constructing a tree from the given input features through a divide and conquer process (DAC). DAC process in the decision tree is based on a recursive partitioning of the input feature spaces into many subspaces according to the value of a candidate input feature. The candidate feature is selected by using a goodness measure of the feature from a set of ranked features. Different goodness measure can be used for ranking including Gain ratio, a likelihood ratio. The leaves of the tree are labeled either as a class value or a probability distribution of the class. Bagging can be performed in order to improve the stability and accuracy of the decision tree [19]–[22].

The classification efficiency of the proposed approach is compared with a Support Vector Machine (SVM) [51] with or without using GANNIGMA hybrid feature classification and

Bagging algorithms. In SVMs, examples are represented as points mapped into a space such that the examples of different classes are divided by a hyperplane and a new example is predicted to a clss based on which side of the hyperplane it falls. SVM determines a hyperplane through a training process.

C. PERFORMANCE MEASURE

The performance of the proposed classification approach is evaluated using different standard measures based on receiver operating characteristics (ROC) graph metrics, True Positive (TP), True negative (TN), False Positive (FP) and False Negative (FN), whereTP is the total number of true positives (patient's with brain tumor), TN is the total number of true negatives; FN is the total number of false negatives, and FP is the total number of false positives. ROC is important performance metric to evaluate the performances of a classifier. In ROC, True positive (TP) rate is plotted on the Y-axis of ROC graph and FP rate is plotted on the X-axis of the graph. The method to calculate area under ROC graph is calculated from the unit square area from ROC and is in the range of 0 to 1.

$$RecallorTurePositiveRate\,(TPR) = \frac{\sum Truepositive}{\sum Conditionpositive} = \frac{TP}{TP+FN} \tag{12}$$

$$FalsePositiveRate\,(FPR) = \frac{\sum Falsepositive}{\sum Conditionnegative} = \frac{FP}{N} = \frac{FP}{FP+TN} \tag{13}$$

$$Precision = \frac{\sum Truepositive}{\sum Testoutcomepositive} = \frac{TP}{TP+FP} \tag{14}$$

The F-measure (F), a measure that combines precision and recall as the harmonic mean of precision and recall is also computed as follows:

$$F = 2 \times \frac{precision \times recall}{precision + recall} \tag{15}$$

IV. RESULT AND DISCUSSION

The proposed approach is tested on a brain tumor dataset of 63 samples with and without the 1p,19q co-deletion. The results are summarized in Table 1 and discussed as follows. Comparisons of the classification accuracy for different combinations of feature selection and classification techniques are summarized in Table 2. Bagging with decision tree performed better compared to other combinations is similar to the findings of [19]. However, the approach [19] is applied for the classification of thyroid disease.

The decision rules obtained by the wrapper approach GANNIGMA+ Bagging + decision Tree are presented in Fig 4(a) which has a tree size of 19. In Fig 4(b), GANNIGMA + MRMR + Decision Tree achieves a smaller tree size of 17. However, while using the hybrid feature selection (GANNIGMA+ MRMR + Bagging+ Decision Tree), the decision rules obtained are more simplified as in Fig 5.

ShapeFactor <= 0.310617: Y
ShapeFactor > 0.310617: N
Tree size: 3

FIGURE 5. Rule from GANNIGMA+ MRMR+ Bagging+ Decision Tree.

The results indicate that GANNIGMA feature selection based proposed ensemble approach provides a more simplified decision rule with higher accuracies, which can be incorporated in neuropathology diagnostics. The imbalanced dataset is an inherent limitation in healthcare data, which is overcome by globally optimized feature selection, bootstrapping and cross-validation. The segmentation of the images and the selection of the morphological features requires further work to improve the classification accuracy [52]–[54].

V. CONCLUSION

The proposed hybrid feature selection with ensemble classification technique in this paper combines a Maximum Relevance and Minimum Redundancy filter heuristic with a globally optimized wrapper heuristic GANNIGMA. The proposed approach aggregates the complementary properties of a filter and a wrapper heuristics and integrates that in the ensemble classification for brain tumor classification. The results clearly indicate that the proposed feature selection and ensemble classification with bagging and decision tree outperform all other existing algorithms and are able to provide a simplified diagnostic rule set that can be used in pathology diagnosis for imbalanced brain tumor dataset. Future work may include application of different search strategies in the feature selection and ensemble techniques with the additional morphological features. Also a statistical approach using regression analysis can be applied to generate pathology diagnostic rule and can be cpompared with the current approach in future.

REFERENCES

[1] M. E. Boon, L. P. Kok, and S. Beck, "Histological validation of neural-network assisted cervical screening: Comparison with the conventional approach," *Cell Vis.*, vol. 2, pp. 23–27, 1995.
[2] K. Chan, T.-W. Lee, P. A. Sample, M. H. Goldbaum, R. N. Weinreb, and T. J. Sejnowski, "Comparison of machine learning and traditional classifiers in glaucoma diagnosis," *IEEE Trans. Biomed. Eng.*, vol. 49, no. 9, pp. 963–974, Sep. 2002.
[3] M. Cruz-Monteagudo, M. N. D. S. Cordeiro, and F. Borges, "Computational chemistry approach for the early detection of drug-induced idiosyncratic liver toxicity," *J. Comput. Chem.*, vol. 29, no. 4, pp. 533–549, Mar. 2008.
[4] D. Delen, A. Oztekin, and Z. Kong, "A machine learning-based approach to prognostic analysis of thoracic transplantations," *Artif. Intell. Med.*, vol. 49, no. 1, pp. 33–42, May 2010.
[5] L. J. Mango, "Computer-assisted cervical cancer screening using neural networks," *Cancer Lett.*, vol. 77, nos. 2–3, pp. 155–162, Mar. 1994.
[6] M. Gadji, D. Fortin, A.-M. Tsanaclis, and R. Drouin, "Is the 1p/19q deletion a diagnostic marker of oligodendrogliomas?" *Cancer Genet. Cytogenetics*, vol. 194, no. 1, pp. 12–22, Oct. 2009.
[7] C. Ramirez *et al.*, "Loss of 1p, 19q, and 10q heterozygosity prospectively predicts prognosis of oligodendroglial tumors—Towards individualized tumor treatment?" *Neurol. Oncol.*, vol. 12, pp. 490–499, Feb. 2010.
[8] J. F. Parkinson *et al.*, "The impact of molecular and clinical factors on patient outcome in oligodendroglioma from 20 years' experience at a single centre," *J. Clin. Neurosci.*, vol. 18, no. 3, pp. 329–333, Mar. 2011.

[9] D. Scheie *et al.*, "Can morphology predict 1p/19q loss in oligodendroglial tumours?" *Histopathology*, vol. 53, no. 5, pp. 578–587, Nov. 2008.
[10] T. Rothhammer and A.-K. Bosserhoff, "Epigenetic events in malignant melanoma," *Pigment Cell Res.*, vol. 20, no. 2, pp. 92–111, Apr. 2008.
[11] J. Chen *et al.*, "WLD: A robust local image descriptor," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 32, no. 9, pp. 1705–1709, Sep. 2010.
[12] E. Ott, *Chaos in Dynamical Systems*. London, U.K.: Cambridge Univ. Press, 1994.
[13] H. F. Jelinek, R. M. Cesar, Jr., J. J. Leandro, and I. Spence, "Automated morphometric analysis of the cat retinal alpha/Y, beta/X and delta ganglion cells using wavelet statistical moment and clustering algorithms," *J. Integr. Neurosci.*, vol. 3, no. 4, pp. 415–432, Dec. 2004.
[14] W. Klonowski, R. Stepien, and P. Stepien, "Simple fractal method of assessment of histological images for application in medical diagnostics," *Nonlinear Biomed. Phys.*, vol. 4, no. 1, pp. 7–15, 2010.
[15] D. Paumgartner, G. Losa, and E. R. Weibel, "Resolution effect on the stereological estimation of surface and volume and its interpretation in terms of fractal dimensions," *J. Microscopy*, vol. 121, no. 1, pp. 51–63, Jan. 1981.
[16] A. Mashiah, O. Wolach, J. Sandbank, O. Uziel, P. Raanani, and M. Lahav, "Lymphoma and leukemia cells possess fractal dimensions that correlate with their biological features," *Acta Haematol.*, vol. 119, no. 3, pp. 142–150, 2008.
[17] R. Abu-Eid and G. Landini, "Morphometrical differences between pseudo-epitheliomatous hyperplasia in granular cell tumours and squamous cell carcinomas," *Histopathology*, vol. 48, no. 4, pp. 407–416, Mar. 2006.
[18] T. Mattfeldt, "Spatial pattern analysis using chaos theory: A nonlinear deterministic approach to the histological texture of tumours," in *Fractals in Biology and Medicine*, vol. 2, G. A. Losa, T. F. Nonnenmacher, D. Merlini, and E. R. Weibel, Eds. Basel, Switzerland: Birkhäuser, 1998, pp. 50–72.
[19] C.-J. Malgorzata, "Boosting, bagging and fixed fusion methods performance for aiding diagnosis," *Biocybern. Biomed. Eng.*, vol. 32, no. 2, pp. 17–31, 2012.
[20] C. A. Lupascu, D. Tegolo, and E. Trucco, "Accurate estimation of retinal vessel width using bagged decision trees and an extended multiresolution Hermite model," *Med. Image Anal.*, vol. 17, no. 8, pp. 1164–1180, Dec. 2013.
[21] H. Parvin, M. MirnabiBaboli, and H. Alinejad-Rokny, "Proposing a classifier ensemble framework based on classifier selection and decision tree," *Eng. Appl. Artif. Intell.*, vol. 37, pp. 34–42, Jan. 2015.
[22] N. Nai-arun and R. Moungmai, "Comparison of classifiers for the risk of diabetes prediction," *Procedia Comput. Sci.*, vol. 69, pp. 132–142, Nov. 2015.
[23] G. J. Postma *et al.*, "On the relevance of automatically selected single-voxel MRS and multimodal MRI and MRSI features for brain tumour differentiation," *Comput. Biol. Med.*, vol. 41, no. 2, pp. 87–97, Feb. 2011.
[24] G. Liu, G. Yan, S. Kuang, and Y. Wang, "Detection of small bowel tumor based on multi-scale curvelet analysis and fractal technology in capsule endoscopy," *Comput. Biol. Med.*, vol. 70, pp. 131–138, Mar. 2016.
[25] P. Bokov, B. Mahut, P. Flaud, and C. Delclaux, "Wheezing recognition algorithm using recordings of respiratory sounds at the mouth in a pediatric population," *Comput. Biol. Med.*, vol. 70, pp. 40–50, Mar. 2016.
[26] D. Sidibé, I. Sadek, and F. Mériaudeau, "Discrimination of retinal images containing bright lesions using sparse coded features and SVM," *Comput. Biol. Med.*, vol. 62, pp. 175–184, Jul. 2015.
[27] T. S. Armstrong, M. Z. Cohen, J. Weinberg, and M. R. Gilbert, "Imaging techniques in neuro-oncology," *Seminars Oncol. Nursing*, vol. 20, no. 4, pp. 231–239, Nov. 2001.
[28] A. Rajendran and R. Dhanasekaran, "Fuzzy clustering and deformable model for tumor segmentation on MRI brain image: A combined approach," in *Proc. Int. Conf. Commun. Technol. Syst. Design*, 2011, pp. 327–333.
[29] D. T. Gering, W. E. L. Grimson, and R. Kikinis, "Recognizing deviations from normalcy for brain tumor segmentation," in *Proc. Int. Conf. Medi. Imag. Comput. Comput.-Assist. Inter. MICCAI 2002. Lecture Notes Comput. Sci.*, vol. 2488, pp. 388–395, Springer, Berlin, 2002.
[30] E. I. Zacharaki *et al.*, "Classification of brain tumor type and grade using MRI texture and shape in a machine learning scheme," *Magn. Reson. Med.*, vol. 62, no. 6, pp. 1609–1618, Dec. 2009.
[31] M. H. Ley *et al.*, "Glial tumor grading and outcome prediction using dynamic spin-echo MR susceptibility mapping compared with conventional contrast-enhanced MR: Confounding effect of elevated rCBV of oligodendrogliomas [corrected]," *Amer. J. Neuroradiol.*, vol. 25, no. 2, pp. 214–221, Feb. 2004.
[32] J. Luts, A. Heerschap, J. A. K. Suykens, and S. Van Huffel, "A combined MRI and MRSI based multiclass system for brain tumour recognition using LS-SVMs with class probabilities and feature selection," *Artif. Intell. Med.*, vol. 40, no. 2, pp. 87–102, Jun. 2007.
[33] R. Cruz-Barbosa and A. Vellido, "SEMI-supervised analysis of human brain tumours from partially labeled MRS information, using manifold learning models," *Int. J. Neural Syst.*, vol. 21, no. 1, pp. 17–29, 2011.
[34] A. Kharrat, K. Gasmi, M. B. Messaoud, N. Benamrane, and M. Abid, "A hybrid approach for automatic classification of brain MRI using genetic algorithm and support vector machine," *Leonardo J. Sci.*, vol. 17, pp. 71–80, Dec. 2010. [Online]. Available: http://ljs.academicdirect.org
[35] D. C. Shubhangi and P. S. Hiremath, "Support vector machine (SVM) classifier for brain tumor detection," in *Proc. Int. Conf. Adv. Comput., Commun. Control*, Jan. 2009, pp. 444–448.
[36] W. E. Reddick, J. O. Glass, E. N. Cook, T. D. Elkin, and R. J. Deaton, "Automated segmentation and classification of multispectral magnetic resonance images of brain using artificial neural networks," *IEEE Trans. Med. Imag.*, vol. 16, no. 6, pp. 911–918, Dec. 1997.
[37] M. C. Clark, L. O. Hall, D. B. Goldgof, R. Velthuizen, F. R. Murtagh, and M. S. Silbiger, "Automatic tumor segmentation using knowledge-based techniques," *IEEE Trans. Med. Imag.*, vol. 17, no. 2, pp. 187–201, Apr. 1998.
[38] B. Scholkopf and A. J. Smola, *Learning With Kernels Support Vector Machines, Regularization, Optimization and Beyond*. Cambridge, MA, USA: MIT Press, 2001.
[39] M. Huml, R. Silye, G. Zauner, S. Hutterer, and K. Schilcher, "Brain tumor classification using AFM in combination with data mining techniques," *BioMed Res. Int.*, vol. 2013, 2013, Art. no. 176519, doi.org/10.1155/2013/176519.
[40] C. Ding and H. Peng, "Minimum redundancy feature selection from microarray gene expression data," *J. Bio. Comput. Biology*, vol. 3, no. 2, pp. 185–205, 2005.
[41] E. I. Papageorgiou *et al.*, "Brain Tumor characterization using the soft computing technique of fuzzy cognitive maps," *Appl. Soft Comput.*, vol. 8, no. 1, pp. 820–828, Jan. 2008.
[42] N. Nabizadeh and M. Kubat, "Brain tumors detection and segmentation in MR images: Gabor wavelet vs. statistical features," *J. Comput. Elect. Eng.*, vol. 45, pp. 286–301, Jul. 2015. [Online]. Available: http://dx.doi.org/10.1016/j.compeleceng.2015.02.007 0045-7906/
[43] K. D. Kharat, P. P. Kulkarni, and M. B. Nagori, "Brain tumor classification using neural network based methods," *Int. J. Comput. Sci. Inform.*, vol. 1, no. 4, pp. 2231–2236, 2012.
[44] Q.-u. Ain, I. Mehmood, S. M. Naqi, M. A. Jaffar, "Bayesian classification using DCT features for brain tumor detection," in *Proc. 14th Int. Conf. Knowl.-Based Intell. Inf. Eng. Syst.*, Heidelberg, Germany, Sep. 2010, pp. 340–349.
[45] M. A. Mammadov, "A new global optimization algorithm based on a dynamical systems approach," in *Proc. Int. Conf. Optim., Tech. Appl. (ICOTA)*, Ballarat, VIC, Australia, 2004, pp. 1–11.
[46] M. Mammadov, A. Rubinov, and J. Yearwood, *Dynamical Systems Described by Relational Elasticities With Applications, Continuous Optimisation: Current Trends and Applications*, 2005, pp. 365–387.
[47] C.-N. Hsu, H.-J. Huang, and S. Dietrich, "The ANNIGMA-wrapper approach to fast feature selection for neural nets," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 32, no. 2, pp. 207–212, Apr. 2002.
[48] S. Huda, J. Yearwood, and A. Strainieri, "Hybrid wrapper-filter approaches for input feature selection using maximum relevance and artificial neural network input gain measurement approximation (ANNIGMA)," in *Proc. 4th Int. Conf. Netw. Syst. Secur.*, Sep. 2010, pp. 442–449.
[49] H. Peng, C. Ding, and F. Long, "Minimum redundancy feature selection from microarray gene expression data," *IEEE Intell. Syst.*, vol. 3, no. 2, pp. 70–71, Germany: Springer, US, 2005.
[50] L. Breiman, "Bagging predictors," *Mach. Learn.*, vol. 24, no. 2, pp. 123–140, Aug. 1996, doi:10.1007/BF00058655.
[51] C. Cortes and V. Vapnik, "Support-vector networks," *Mach. Learn.*, vol. 20, no. 3, pp. 273–297, Sep. 1995, doi: 10.1007/BF00994018.

[52] H. Jelinek, S. Matthews, P. Succar, C. S. McLachlan, and M. Buckland, "Diagnosis of oligodendroglioma: Molecular and classical histological assessment in the twenty-first century," *Asia–Pacific J. Clin. Oncol.*, vol. 8, no. 3, pp. 213–216, Sep. 2012, doi: 10.1111/j.1743-7563.2012.01527.x.
[53] S. Matthews, P. Succar, H. F. Jelinek, B. McParland, M. Buckland, and C. S. McLachlan, "Establishing a reference range for oligodendroglioma classification using Higuchi dimension analysis," in *Proc. 9th IASTED Int. Conf. Biomed. Eng., BioMed*, 2012, pp. 25–28.
[54] L.-T. Kuo *et al.*, "Genetic and epigenetic alterations in primary-progressive paired oligodendroglial tumors," *PLoS ONE*, vol. 8, no. 6, p. e67139, 2013.
[55] G. Aloi *et al.*, "Software defined radar: Synchronization issues and practical implementation," presented at the ACM Int. Conf. Series, Oct. 2011, Art. no. 48, doi: 10.1145/2093256.2093304.
[56] G. Fortino, R. Giannantonio, R. Gravina, P. Kuryloski, and R. Jafari, "Enabling effective programming and flexible management of efficient body sensor network applications," *IEEE Trans. HumanŰMach. Syst.*, vol. 43, no. 1, pp. 115–133, Jan. 2013.
[57] S. Iyengar, F. T. Bonda, R. Gravina, A. Guerrieri, G. Fortino, and A. Sangiovanni-Vincentelli, "A framework for creating healthcare monitoring applications using wireless body sensor networks," in *Proc. 3rd Int. Conf. Body Area Netw., BodyNets (ICST)*, Mar. 2008, p. 8.
[58] G. Fortino and V. Giampà, "PPG-based methods for non invasive and continuous blood pressure measurement: An overview and development issues in body sensor networks," in *Proc. IEEE Int. Workshop Med. Meas. Appl. (MeMeA)*, Apr. 2010, pp. 10–13.
[59] G. Fortino, S. Galzarano, R. Gravina, and W. Li, "A framework for collaborative computing and multi-sensor data fusion in body sensor networks," *Inf. Fusion*, vol. 22, pp. 50–70, Mar. 2015.

SHAMSUL HUDA was a Lecturer with Federation University, Australia. He was an Assistant Professor with the Computer Science Department, Khulna University of Engineering and Technology, Bangladesh. He is currently a Lecturer with the School of Information Technology, Deakin University, Australia. He has authored over 50 journal and conference papers in well reputed journals, including the IEEE Transactions. His main research area is computational intelligence, information security, optimization approaches to data mining, health informatics.

JOHN YEARWOOD is currently the Head of the School of Information Technology, Deakin University, Australia. His main research areas are machine learning, optimization, and information security. He has authored two books and over 200 refereed journal, book chapter, and conference articles. He was the Editor-in-Chief of the *Journal of Research and Practice in Information Technology*, and a Reviewer for many journals.

HERBERT F. JELINEK (M'10) received the B.Sc. degree(Hons.) in human genetics from the University of New South Wales, Sydney, Australia, the Graduate Diploma degree in neuroscience from the Australian National University, Canberra, Australia, and the Ph.D. degree in medicine from the University of Sydney, Australia. He is currently an Honorary Clinical Associate Professor with the Australian School of Advanced Medicine, Macquarie University, Sydney, and a member of the Center for Research in Complex Systems, Charles Sturt University, Albury, Australia. He has been an organizing a rural diabetes complications screening research project for over ten years in Australia and has authored widely in ECG signal processing, diabetic retinopathy image analysis, data mining, and biomarkers associated with diabetes disease progression. His current research interests include neurogenetics of diabetes and cognitive function. He is a member of the IEEE Biomedical Engineering Society and the Australian Diabetes Association.

MOHAMMAD MEHEDI HASSAN (M'12) received the Ph.D. degree in Computer Engineering from Kyung Hee University, South Korea in February 2011. He is currently an Assistant Professor of Information Systems Department with the College of Computer and Information Sciences (CCIS), King Saud University (KSU), Riyadh, Saudi Arabia. He was a Post-Doctoral Fellow at Computer Engineering Department, Kyung Hee University, South Korea from March, 2011 to November, 2011. He is one of the founding members of Chair of Pervasive and Mobile Computing (CPMC) at CCIS, KSU and successfully managing its research program, which transformed the chair as one of the best chairs of research excellence in the college. He was a recipient of the Best Paper Award from the CloudComp 2014 Conference, China, in 2014, the Excellence in Research Award from the CCIS, KSU, in 2015 and 2016. He has authored over 100 research papers in the journals and conferences of international repute such as the *IEEE Wireless Communication Magazine*, the *IEEE Network Magazine*, the *IEEE Communication Magazine*, the IEEE Transaction on Computers, the IEEE Transaction on Services Computing, the *Future Generation Computing Systems*, and the ACM Multimedia conference. He has served as the Chair and a Technical Program Committee member in numerous international conferences/workshops, including the IEEE CCNC, the ACM BodyNets, and the IEEE HPCC. He was also the Guest Editor of several international ISI-indexed journals. He has secured several national and international research grants in the domain of cloud computing and sensor network.

GIANCARLO FORTINO is currently an Associate Professor of Computer Science with the Dipartiment of di Informatica, Elettronica e Sistemistica, University of Calabria. His research is mainly focused on methodologies, frameworks and tools for programming distributed computing systems, distributed health analytics, and cloud based health analytics.

MICHAEL BUCKLAND received the MBBS degree from The University of Sydney and the Ph.D. degree in neurobiology from the Garvan Institute of Medical Research, and the specialist training was undertaken with Royal Prince Alfred (RPA) and Royal North Shore Hospitals, Sydney, Australia. He is currently a Senior Staff Specialist Neuropathologist, the Head of the Department of Neuropathology with RPA Hospital, the Head of the Molecular Neuropathology Program with the Brain and Mind Research Institute, The University of Sydney, and a Co-Director of the Multiple Sclerosis Research Australia Brain Bank.