IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received January 2, 2019, accepted February 6, 2019, date of publication February 12, 2019, date of current version April 3, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2898897

A Novel Multi-Stage Fusion based Approach for Gene Expression Profiling in Non-Small Cell Lung Cancer

MUHAMED WAEL FAROUQ[1,2], WADII BOULILA[3,4], (Senior Member, IEEE), MEDHAT ABDEL-AAL[1], AMIR HUSSAIN[2,5], (Senior Member, IEEE), AND ABDEL-BADEEH SALEM[6]

[1]Department of Statistics, Mathematics and Insurance, Ain Shams University, Cairo 11566, Egypt
[2]School of Computing, Edinburgh Napier University, Edinburgh EH10 5DT, U.K.
[3]IS Department, College of Computer Science and Engineering, Taibah University, Medina 41411, Saudi Arabia
[4]RIADI Laboratory, National School of Computer Sciences, University of Manouba, Manouba 2010, Tunisia
[5]Taibah Valley, Taibah University, Medina 42353, Saudi Arabia
[6]Faculty of Computer Science, Ain Shams University, Cairo 11566, Egypt

Corresponding author: Wadii Boulila (wadii.boulila@riadi.rnu.tn)

ABSTRACT **Background:** Non-small cell lung cancer is defined at the molecular level by mutations and alterations to oncogenes, including AKT1, ALK, BRAF, EGFR, HER2, KRAS, MEK1, MET, NRAS, PIK3CA, RET, and ROS1. A better understanding of non-small cell lung cancer requires a thorough consideration of these oncogenes. However, the complexity of the problem arises from high-dimensional gene vector space, which complicates the identification of cluster boundaries, and hence gene expression cluster membership. This paper aims to analyze potential biological biomarkers for tumorigenesis in lung cancer based on different treatment solutions. **Results:** Genes BRAF, RET, and ROS1 show an overexpression transition by one cluster from non-treatment to treatment states, followed by a stabilization in the 3 treatment states at the same cluster. Genes MET, ALK, and PIK3CA show an overexpression transition by two clusters from non-treatment to treatment states, followed by a stabilization in the 3 treatment states at the same cluster. SME1 shows an under-expression transition by two clusters from non-treatment to the treatment states, a stabilization in the 3 treatment states at the same cluster. **Conclusions:** We present a novel fusion-based approach for gene expression profiling of non-small cell lung cancer under non-thermal plasma treatment. The main contribution of the proposed approach is to exploit Dempster–Shafer evidence theory-based data fusion to combine information from different samples in the considered dataset. This minimizes uncertainty and enhances the reliability and validity of decisions, leading to a better description of genes related to non-small cell lung cancer. We also propose use of fuzzy c-means-with-range clustering to track changes of genes' states under different non-thermal plasma treatments.

INDEX TERMS Gene expression, Dempster Shafer, evidence theory, data fusion, clustering, non-small cell lung cancer.

I. INTRODUCTION

DNA microarray has made the analysis of the dynamics and interactions of thousands of genes simultaneously possible. The inference of expression data is guided by the following facts: 1) expression data are high dimensional and complex, and 2) dynamic relations exist among thousands of genes simultaneously and/or sequentially. Dynamic relations on one hand may reveal cascade interactions between genes. Indeed, the expression of one gene may alter the transcription rate of another one. On the other hand, dynamic relations may show coherent patterns (i.e. genes with similar expressions suggest they are more likely co-regulating each other or to be regulated by a parental gene). Expression data may also show both cascade interactions and coherent patterns. Further complicating the inference is the fact that sample profile-to-gene profile ratio is typically very small.

Expression data can roughly be visualized as individual cells or expression profiles. An individual cell in the gene expression matrix represents the expression levels of each gene under each sample or time point. An alternative to the

The associate editor coordinating the review of this manuscript and approving it for publication was Bora Onat.

individual cell representation of gene expression data is expression profile representation. Following the expression matrix structure and the fact that sample-to-gene ratio is usually small, gene expression profile can be thought of as a vector represented in the samples' vector space. In addition, the sample expression profile can be thought of as a vector represented in the high-dimensional gene vector space.

Tumor suppressor gene (TSG) protects the cell from cancer by controlling the cell cycle and promoting apoptosis through a number of mechanisms. First, TSG has a repressive effect on genes that dutifully controls cell cycle; therefore, this repressive effect on the gene inhibits cell division. Second, TSG tethers the DNA damage to the cell cycle, such that if DNA damage occurs and cannot be repaired, the cell initiates apoptosis.

The main goal of this paper is to analyze potential biological biomarkers for tumorigenesis non-small cell lung cancer based on different treatment solutions. The identification of cluster boundaries and the membership of gene expression clusters is known to be a difficult task due to the high dimensionality of the gene vector space. We thus present a novel fusion-based approach for gene expression profiling of non-thermal plasma treatment of non-small cell lung cancer. The proposed approach is based on data fusion and fuzzy clustering to analyze genes' clustering. The context of our work is molecular expression profiling of biological biomarkers for non-small cell lung cancer.

The remainder of this paper is organized as follows. In Section 2, we present related works. Section 3 describes the considered dataset. In section 4, we detail the proposed framework. Simulation results are presented in Section 5. Finally, Section 6 concludes the paper and outlines some future work directions.

II. RELATED WORKS

Clustering in the gene expression context helps to elucidate gene functions and reveal tumor typology [1]. The aim of gene-based clustering techniques is the projection of high-dimensional individual clusters to an optimal reduced dimension of group clusters, to determine distinct gene expression levels that can aid in the understanding of gene functions. Sample-based clustering techniques search for samples with similar expression patterns to specify different (distinct) tumor types. Due to the small sample-to-gene ratio, sample-based clustering is a challenging task [2]. In the literature, several works attempt to apply clustering techniques in the gene expression context.

Microarray technology has made possible the measurement of expression of thousands of genes simultaneously. The process of identifying the number of informative genes is crucial and is an area of intensive research. For example, Fisher discriminant criterion (FDC), Cross projection (CP) and discrete partition are measures studied by Cao and Zhu [3] to identify genes that are predictive of clinical conditions. However, the imprecision of state-of-the-art classification methods and uncertainties in the microarray data are problems that need to be addressed. In their work, the authors introduced a knowledge-based belief reasoning system (BRS) to solve the problem of uncertainties in classification results. This initially sorts the results of FDC, CP and DP of gene values separately and picks up the top 5% of genes. Then, Dempster-Shafer evidence fusion is performed on the candidate list to return the fusion genes.

A common challenge in hierarchal clustering is the determination of the cut-off level or splitting level of the dendrogram. The commonly fixed height branch split-level is inflexible. Langfelder *et al.* [4] introduced flexible dendrogram branch cutting methods, named the dynamic tree cut package, which is available online. By combining hierarchical clustering and partitioning-around medoid advantages, dynamic cutting methods give better outlier detection. The use of these methods was illustrated on simulated gene expression data.

Richards *et al.* [5] compared results of four clustering algorithms namely, CRC, k-means, ISA, and memISA on datasets of microarray brain expression. The comparison is based on three performance measures: speed, gene coverage, and GO-enrichments. Although ISA and memISA GO-enrichments slightly outperform k-means enrichments but this is at the cost of gene coverage and speed. The authors report that k-means outperforms the other three algorithms with a gene coverage of 100%. However, combining k-means and ISA or memISA can further improve the clustering performance. The estimation of number of clusters for the PAM clustering algorithm is crucial.

Wang *et al.* [6] addressed the issue of accurate estimation of number of clusters in the PAM algorithm by introducing a system evolution method. The authors proposed to analyze cluster structures of a dataset from the viewpoint of a pseudo thermodynamics system by using partitioning and merging processes. The experimental results of gene expression demonstrate a good performance of the introduced system on data structures (i.e. data structures are well separated or data present a slightly overlapping structure).

Zhang and Sun [7] exploited evidence theory as one of the mathematical theories to construct more informative gene regulatory networks (GRNs) by combining multiple biological data sources. Specifically, they proposed to fuse transcription factor and gene expression data using Dempster-Shafer evidence theory. In addition, a smooth probabilistic model was applied to the transcription factor data while the Pearson correlation model was applied to gene expression data. Finally, a dynamic Bayesian network algorithm was used for learn, as part of the proposed system.

Krejnik and Klema [8] proposed a methodology to impartially verify the applicability of particular types of gene clustering approaches. The verification was conducted as part of a predictive classification framework and focussed on prior biological knowledge-based functional clustering. The analysis was performed in terms of gene expression classification and used predictive accuracy as an unbiased performance measure. Features of biological samples that originally corresponded to genes were replaced by features that represented centroids of gene clusters, and then used for classifier learning. The authors validated their approach using 10 benchmark datasets and compared the performance of their approach with existing clustering methods.

Chen and Jian [9] presented a segmentation method for clustering gene expression data based on graph-regularized subspace. The goal of their approach was to combine graph regularization with subspace segmentation, for modeling the intrinsic geometrical structure of the data space. Specifically, the authors used the Sylvester equation to find a global optimal solution for segmentation of the graph regularized subspace. The proposed approach was evaluated using eight gene expression datasets and compared with subspace segmentation methods, traditional clustering methods, and clustering methods based on nonnegative matrix factorization.

Jiang *et al.* [10] proposed fuzzy c-means (FCM) clustering based on weights and gene expression programming to improve the performance of classical FCM. Specifically, the authors first introduced a similarity calculation formula to obtain the average entropy of data. Next, cluster centers were computed through gene expression programming by encoding them as chromosomes, in order to determine the appropriate cluster centers. The approach was validated based on ten UCI datasets and compared to classical FCM results.

Dutta and Saha [11] explored the use of multi-objective optimization based on genetic clustering techniques. The objective was to classify genes into groups with respect to their functional similarities and biological relevance. Specifically, the authors developed a quality measure to compute the goodness of gene-clusters, namely confidence score of protein-protein interaction. Further, they proposed multi-objective based clustering which employed integrated information of expression values of microarray dataset and protein-confidence score for protein interactions, in order to select both statistically and biologically relevant genes. Experiments were performed on three datasets of real-life gene expression and results compared to existing techniques.

Paul and Shill [12] proposed annotations for gene ontology based on a semi-supervised clustering algorithm. The developed algorithm was termed GO fuzzy relational clustering, in which one gene could be assigned to multiple clusters. The algorithm utilised biological knowledge, available in the form of a gene ontology, as prior knowledge, along with the gene expression data. The prior knowledge was found to help improve the coherence of the groups. The algorithm was tested using two yeast datasets and results compared with other state-of-the-art clustering algorithms.

III. DATASET

Plasma is one of the fundamental states of matter. It has gas property (i.e. no definite shape or volume) and unlike solid and liquid, is less dense [13]. It is created by a process called ionization, wherein atoms or molecules of a gas acquire a negative or positive charge by heating, or are subjected to a strong electromagnetic field at relatively very high temperature. This process causes a gain or loss of electrons, which leads to forming positively or negatively charged particles called ions. Plasma can be thermal or non-thermal. Thermal plasma has the same temperature for electrons, ions and neutrals. While in non-thermal plasma, the temperature of electrons is higher than that of ions and neutrals.

TABLE 1. Dataset notations.

NT	Non-treatment or control group
SE	Short exposure non-thermal plasma treatment, measured post 4 hours of treatment
LE post 1hr	Long exposure non-thermal plasma treatment, measured post 1 hour of treatment
LE post 2hr	Long exposure non-thermal plasma treatment, measured post 2 hours of treatment
LE post 4hr	Long exposure non-thermal plasma treatment, measured post 4 hours of treatment

Recent technological advances have made the use of non-thermal plasma in the medical field a reality. Cancer arises as a disorder of one of the organs' cell function, specifically cell division. This causes an abnormal growth of cells and can spread from one organ to another through a process called metastases. Hou *et al.* [14] provided tumor cellular gene expression profile of lung adenocarcinoma, upon treatment with non-thermal atmospheric plasma. Their data provided the transcriptome of the tumor cell prior to treatment for three cases (which we term samples in our work), as well as the transcriptome upon short exposure, non-thermal plasma treatment and long exposure, non-thermal plasma treatment for each of the three samples. The transcriptome was obtained at different time points. Specifically, the short exposure for non-thermal plasma treatment transcriptome of the tumor cell was measured post 4 hours, while the long exposure for non-thermal plasma treatment transcriptome was measured post 1, 2 and 4 hours. This data is publicly accessible through the NCBI Gene Expression Omnibus under GEO accession number GSE59997.

The heat map in Figure 1 is constructed based on gene expression analysis and gives an overview of the considered dataset. In the columns, we depict 15 dataset samples (NT1 refers to NT sample 1, NT2 refers to NT sample 2, NT3 refers to NT sample 3, SE1 refers to SE sample 1, SE2 refers to SE sample 2, SE3 refers to SE sample 3, LE1hr1 refers to LE post 1hr sample 1, LE1hr2 refers to LE post 1hr sample 2, LE1hr3 refers to LE post 1hr sample 3, LE2hr1 refers to LE post 2hr sample 1, LE2hr2 refers to LE post 2hr sample 2, LE2hr3 refers to LE post 2hr sample 3, LE4hr1 refers to LE post 4hr sample 1, LE4hr2 refers to LE post 4hr sample 2 and LE4hr3 refers to LE post 4hr sample 3). In the rows, we depict values of genes (49395 genes in our dataset).

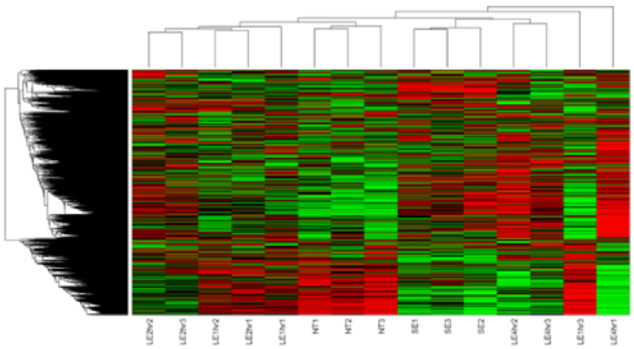

FIGURE 1. Heat map of the considered dataset.

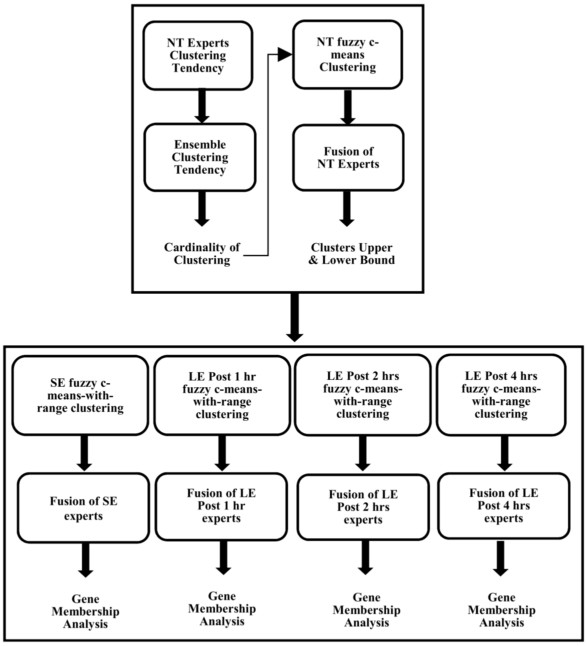
NT Experts Clustering Tendency
Ensemble Clustering Tendency
Cardinality of Clustering
NT fuzzy c-means Clustering
Fusion of NT Experts
Clusters Upper & Lower Bound
SE fuzzy c-means-with-range clustering
LE Post 1 hr fuzzy c-means-with-range clustering
LE Post 2 hrs fuzzy c-means-with-range clustering
LE Post 4 hrs fuzzy c-means-with-range clustering
Fusion of SE experts
Fusion of LE Post 1 hr experts
Fusion of LE Post 2 hrs experts
Fusion of LE Post 4 hrs experts
Gene Membership Analysis
Gene Membership Analysis
Gene Membership Analysis
Gene Membership Analysis

FIGURE 2. Proposed framework.

IV. PROPOSED FRAMEWORK

Our proposed microarray framework based on gene expression fusion is presented in Figure 2. Individual stages in the framework are described in Algorithm 1.

Next, in this section, and the next, we discuss details of the framework showing in Figure 2, and Algorithm 1. As with most clustering problems, we start by probing the clustering tendency. This step aims to determine if the considered data exhibits an intrinsic predisposition to cluster into distinct groups. This step is performed for the non-treatment samples. Here, we employ three techniques to measure the clustering tendency. Two statistical techniques namely Hopkins [15] and Cox-Lewis [16], and one visual technique namely Visual Assessment of cluster Tendency (VAT) [17]. Since we have three samples for the NT dataset, an ensemble-learning step is needed to reach a hybrid decision on the cluster tendency of this dataset. In the case of presence of a clustering structure in our dataset, a further step is required to determine the clustering cardinality.

The third step in our proposed approach is to perform a fuzzy c-means clustering for the three NT samples [18], [19]. The main goal of this step is to compute masses needed in the next fusion step. In this paper, the squared Euclidean distance metric is used for fuzzy c-means clustering. The fusion is based on evidence theory and allows combining the three NT samples into a single clustering output, that gives a better description of the NT dataset.

The last step performed for NT samples is determination of upper and lower bounds for each cluster, obtained after the fusion process. These bounds are then used for clustering of the SE and LE samples. The SE treatment aims to determine changes in gene membership from NT and SE samples. Therefore, the upper and lower bounds for each cluster in the SE samples are taken to be the same for NT samples. Next, we perform fuzzy c-means clustering while considering this constraint. Evidence-then based fusion is then applied to the results of clustering the three SE samples. The final step of the SE treatment is to determine genes that preserved their clusters in the NT and SE treatments.

The same process applied to the SE dataset is repeated for the datasets LE post 1hr, post 2hr and post 4hr.

In this work, we are interested in discovering genes related to non-small cell lung cancer. In the considered dataset, we have two different non-thermal plasma treatments with two doses: short exposure and long exposure. The goal is to analyze the effect of different treatment strategies on the non-small cell lung cancer genes. For each of the 5 states we have (NT, SE, LE post 1hr, LE post 2hr and LE post 4hr), three cases are present. The idea is to combine information from three cases in each state into a single decision for each state. This decision will depict a better description of genes related to non-small cell lung cancer. The process of data fusion will ensure data integration from three cases to produce more consistent and accurate information on genes related to non-small cell lung cancer, than that provided by any individual case.

The next section describes the main steps of the fusion process and the fuzzy c-means-with-range clustering.

A. EVIDENCE FUSION PROCESS

A key contribution of this paper is to exploit advantages of redundancies and complementarities between information obtained by different data of gene-expression samples to propose more accurate and relevant decisions on the data. This is ensured by employing evidence-fusion theory which aims to address the uncertainty related to subjective judgements provided by different data samples of the same treatment (NT, SE and LE), to disregard incompatible and conflicting opinions, and to incorporate information using belief structures. This will provide reliability and validity for resulting decisions on gene expression data.

In this paper, we propose to fuse samples data using the Dempster-Shafer method [20]–[23]. Let D (frame of

Algorithm 1 FGEP (Fusion of Samples of Gene Expression Profiling) Steps

```
1.Clustering tendency and cardinality
    1.1 Clustering tendency based on control group
        (Non-treatment) samples.
    1.2 Ensemble learning for clustering tendency.
    1.3 Clustering cardinality based on control group
        (Non-treatment) samples.
2. Non-treatment
    2.1 Apply fuzzy c-means clustering for each sample.
    2.2 Fusion of 3 NT samples.
    2.3 After fusion for each cluster, we determine the upper
        and lower bounds.
3. SE treatment
    3.1 Apply a fuzzy c-means-with-range clustering
        for each sample of SE.
    3.2 Fusion of 3 SE samples.
    3.3 Test the results (changes in gene membership)
        between NT fusion and SE fusion.
4. LE treatment post 1hr
    4.1 Apply LE post 1hr fuzzy c-means-with-range clustering for each sample.
    4.2 Fusion of 3 samples of LE post 1hr.
    4.3 Test the results (changes in gene membership) between NT fusion and LE post 1hr fusion.
5. LE Treatment post 2hr
    5.1 Apply fuzzy c-means-with-range clustering
        for each sample of LE post 2hr.
    5.2 Fusion of 3 samples of LE post 2hr.
    5.3 Test the results (changes in gene membership) between NT fusion and LE post 2hr fusion.
6. LE Treatment post 4hr
    6.1 Apply a fuzzy c-means-with-range clustering
        for each sample of LE post 4hr.
    6.2 Fusion of 3 samples of LE Post 4hr.
    6.3 Test the results (changes in gene membership)
        between NT fusion and LE Post 4hr fusion.
```

discernment) be a finite set. A mass function on D, where $\mathrm{D} = \{\mathrm{C}_1, \mathrm{C}_2, \ldots, \mathrm{C}_\mathrm{n}\}$, C_i is the cluster i, and n is the number of clusters, is a function m: $2^\mathrm{D} \rightarrow [0,1]$.

$$\sum_{A \subseteq D} m(A) = 1$$

The subsets A of D such that $m(\mathrm{A}) > 0$ are called the focal elements of m. In our case, a mass function m is used to model samples' beliefs about a gene X.

To each mass m can be associated a belief (Bel) and plausibility (Pls) functions defined as follows [24], [25]:

$$Bel\,(A) = \sum_{B \subseteq A, B \neq 0} m\,(B) \forall A \in 2^D$$

$$Pls\,(A) = \sum_{B \cap A \neq 0} m\,(B) = 1 - Bel\left(A^C\right) \forall A \in 2^D$$

In this method, mass functions are combined using Dempster's orthogonal rule [24], [26]

$$m\,(A) = (m_1 \oplus m_2 \oplus, \ldots, m_l \oplus)\,(A) = \frac{\sum_{B_1 \cap, \ldots, \cap B_l = A} m_1\,(B_1)\, m_2\,(B_2)\,, \ldots, m_l\,(B_l)}{1 - K}$$

$$K = \sum_{B_1 \cap, \ldots, B_l \neq 0} m_1\,(B_1)\, m_2\,(B_2)\,, \ldots, m_l\,(B_l)$$

K represents the degree of conflict between samples.

The decision can be taken using one of several rules [25], [26]:

- The maximum of plausibility is defined as

$$x \in C_i \; si \; Bel\,(C_i)\,(x) = \max\{PLs\,(C_k)\,(x)\} 1 \leq k \leq n$$

- The maximum of belief is defined as

$$x \in C_i \; si \; Bel\,(C_i)\,(x) = \max\{Bel\,(C_k)\,(x)\} 1 \leq k \leq n$$

The novel contribution presented here, is feeding the cardinality and the range for each cluster from fusion of NT samples to SE and LE clustering. The next step is application of fuzzy c-means-with-range clustering for each sample data as outined below.

B. FUZZY C-MEANS-WITH-RANGE CLUSTERING

The main goal of fuzzy c-means-with-range clustering is to perform clustering on the SE and LE data while considering the range of each cluster obtained in the previous fusion step.

The proposed process consists of taking clusters resulting from the fusion of NT samples. Then, we compute, for each obtained cluster, the lower and upper bounds for genes belonging to the NT data that are part of the cluster. The next step is to apply Dempster-Shafer based fusion to the resulting fuzzy c-means-with-range clustering.

The same algorithmic steps for SE treatment are applied to the LE treatment with its three states (post 1hr, post 2hrs, and post 4hrs).

The final phase is to explore the presence of change in the cluster size among the NT, SE and LE conditions. This will acts as an indicator of the migration of part/all genes from one cluster to another. Analyzing this migration and the change in each individual gene membership score is used to reveal the effect of non-thermal plasma treatment on tumor transcriptome.

In summary, our proposed scheme, depicted in Figure 3 and Table 2, presents the clustering structure and gene membership of the non-thermal plasma treatment dataset. The aim is to represent the gene dynamics under the three conditions; starting with an initial clustering structure of the non-treatment condition, that is used as an indicator of the cardinality of clustering solutions, and cluster range for the conditions. After determining the cardinality and range

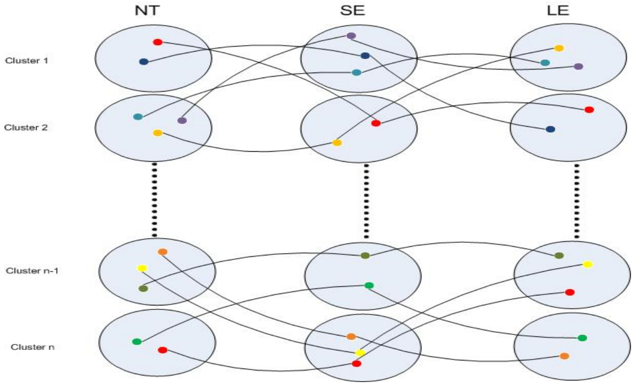

FIGURE 3. Fusion clustering scheme.

TABLE 2. Clustering boundaries.

Clusters	*Fusion NT*	*Fusion SE*	*Fusion LE*
1	a_1 - b_1	a_1-b_1	a_1-b_1
2	a_2 -b_2	a_2 -b_2	a_2 -b_2
.	.	.	.
.	.	.	.
n-1	a_{n-1} -b_{n-1}	a_{n-1} -b_{n-1}	a_{n-1} -b_{n-1}
n	a_n -b_n	a_n -b_n	a_n -b_n

of clusters, we apply the clustering algorithm to reveal the clustering structure or solution for both SE and LE. The result is three clustering structures with the same cardinality and range. Finally, we look for any consensus among the clustering solutions of NT, SE and LE for gene cluster membership.

V. EXPERIMENTS

In this section, a set of experiments are conducted to validate our proposed approach. We start by examining the cluster tendency of the dataset and compute the optimal number of clusters. Next, fuzzy clustering with range and fusion are applied to SE, LE post 1hr, LE post 2hr, and LE post 4hrs. The goal of this step is to determine stable/unstable genes. Finally, an interpretation of the obtained results is presented. The experiments were carried out using Matlab R2016b and performed on a laptop computer with an Intel Core i7-6600U CPU @ 2.60 GHz-2.40GHz (4CPUs), 8GB of RAM, 512GB SSD, and a Windows 10 OS.

A. CLUSTERING TENDENCY AND CARDINALITY

The first step in the proposed approach is examining the cluster tendency of the NT condition of non-thermal plasma treatment data. We used the two statistical indices (Hopkins and Cox Lewis), and visual technique VAT for the three NT samples. Table 3 depicts values of Hopkins and Cox Lewis indices for the three NT samples. Figure 4 shows output images of the VAT algorithm for the three NT samples.Cluster tendency values depicted in Table 3 and results of the VAT algorithm convey a plausible evidence of the presence of clustering structure in the NT data.

TABLE 3. Clustering tendency indices.

	NT		
	Sample 1	*Sample 2*	*Sample 3*
Hopkins	0.91	0.98	0.62
Cox Lewis	1.34	1.48	1.2

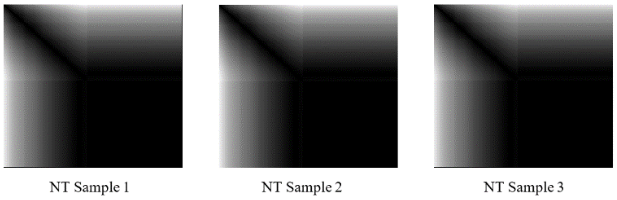

FIGURE 4. VAT images for the NT samples.

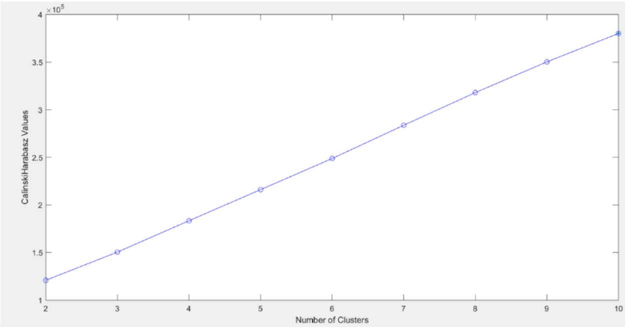

FIGURE 5. Optimal number of clusters given by the Calinski method.

The next step is to determine the number of clusters. Since we do not have any prespecified structure to compare with, we used only internal validation indices to determine cluster cardinality. The cluster cardinality is determined by a combinatorial decision from the three NT samples. We present below the internal validation indices for NT sample 1. The same process is repeated for NT samples 2 and 3. We used three indices: Calinski, Davies Bouldin and Silhouette under different clustering schemes and searched for the optimal number of clusters. The optimal number of clusters for Calinski was found to be 10 with Calinski value of 3.75, as shown in Figure 5. The Davis Bouldin index achieves a minimum value slightly above 0.51 at 8 clusters as illustrated in Figure 6. Figure 7 shows the optimal silhouette value at 2 clusters where it is at a maximum. An average silhouette width close to one indicates a strong cluster structure; the silhouette value at 2 clusters is 0.78.

In order to achieve a robust cluster cardinality combining results from three above indices, we propose an ensemble of internal validity index. Specifically, to use the three internal validity indices, we employ the following equation for normalization of values

$$normalized\ index = \frac{index - \min(index)}{\max(index) - \min(index)}$$

Figure 8 presents the optimal number of clusters, that is 4, corresponding to the highest index value slightly below 0.6.

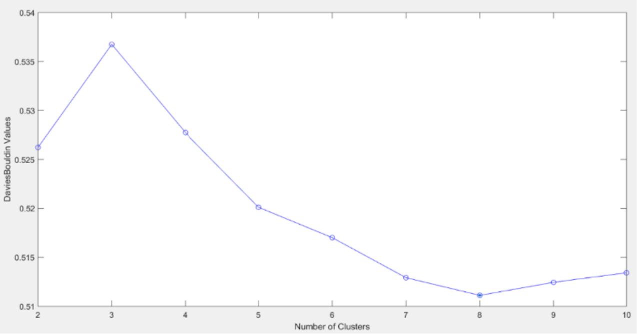

FIGURE 6. Optimal number of clusters given by the Davies Bouldin method.

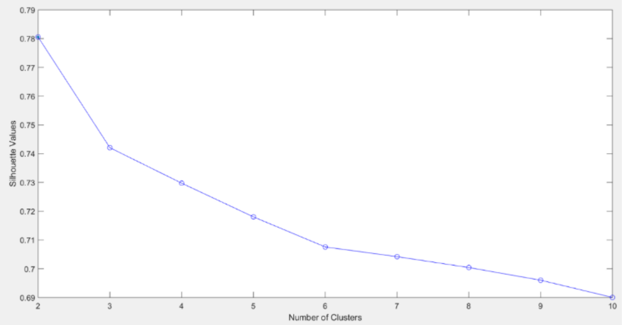

FIGURE 7. Optimal number of clusters given by the Silhouette method.

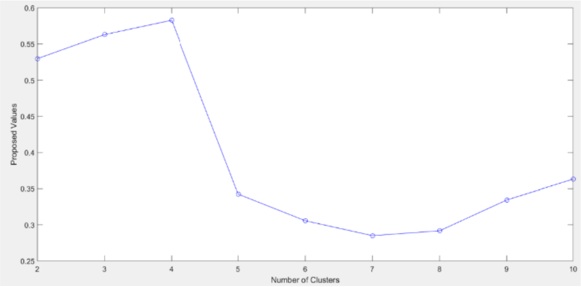

FIGURE 8. Proposed optimal number of clusters.

The same process is repeated for NT samples 2 and 3. The optimal number of clusters for the clustering solution is found to be 4.

B. NT DATA CLUSTERING AND EVIDENCE FUSION

The first step applies fuzzy c-means clustering to each NT sample. Table 4 depicts the range of the 4 clusters and the number of genes belonging to each cluster. As can be seen from Table 1, C1 is bounded between 2.7 to slightly above 5, C2 is bounded between 5 and 7.2, C3 is bounded between 7.2 to slightly above 9.7 and C4 is bounded between 9.7 and 15.1. Moreover, C1 can be seen to be the biggest cluster for the three NT samples, comprising 19011 genes for NT sample 1, 15888 genes for NT sample 2 and 19080 genes for NT sample 3. Whereas, the smallest cluster is C4 comprising 5309 genes for NT sample 1, 5180 genes for NT sample 2 and 5365 genes for NT sample 3. Results of the three NT samples are combined using evidence theory fusion technique, with fusion results depicted in the last column in Table 4. The fusion of NT samples can be seen to sustain the cluster boundaries where C1 is still bounded between 2.7635 to 5.0882 with the biggest cluster comprising 17000 genes, C2 is the second biggest cluster with 140369 genes and bounded between 5.0904 and 7.2614, C3 has 13867 genes with expression profiles between 7.2618 and 9.9480, and finally C4 is the smallest cluster of 4489 genes with expression values bounded between 9.9490 and 15.0773.

TABLE 4. Cluster boundaries and number of genes for the NT group.

	NT sample 1		*NT sample 2*		*NT sample 3*		*Fusion NT*	
	R	***g***	***R***	***g***	***R***	***g***	***R***	***g***
C1	*2.7392*	*19011*	*2.7365*	*158* *88*	*2.7984*	*19080*	*2.7635*	*17000*
	5.0842		*5.1032*		*5.0892*		*5.0882*	
C2	*5.0846*	*13999*	*5.1033*	*141* *38*	*5.0896*	*13821*	*5.0904*	*14039*
	7.2665		*7.2575*		*7.2636*		*7.2614*	
C3	*7.2670*	*11076*	*7.2576*	*111* *89*	*7.2639*	*11129*	*7.2618*	*13867*
	9.8090		*9.7851*		*9.7898*		*9.9480*	
C4	*9.8102*	*5309*	*9.7859*	*518* *0*	*9.7901*	*5365*	*9.9490*	*4489*
	15.1854		*15.2003*		*15.108*		*15.0773*	

TABLE 5. Cluster solution and number of genes for fusion of samples of LE post 1hr, LE post 2hr and LE post 4hr.

		Fusion of LE post 1hr	*Fusion of LE post 2hr*	*Fusion of LE post 4hr*
C 1	*2.7635*	*12907*	*12912*	*12907*
	5.0882			
C 2	*5.0904*	*11824*	*11843*	*11855*
	7.2614			
C 3	*7.2618*	*11868*	*11872*	*11850*
	9.9480			
C 4	*9.9490*	*12796*	*12768*	*12783*
	15.0773			

C. GENERALIZATION OF CLUSTERING AND FUSION

The cluster boundaries after the fusion of NT samples now act as a prespecified clustering structure for the SE post 4hrs, LE post 1hr, LE post 2hrs and LE post 4hrs. Next, we perform clustering for each sample while considering the boundaries of each cluster that were obtained above. Results of fusion of the three samples of LE post 1hr are presented in Table 5. In the same manner, the fusion of three LE post 2hrs and three LE post 4hrs samples are illustrated in Table 5. The fusion results reveal a very close cluster gene size under all 3 LE states, with C1 ranging from 12907 to 12912 genes, C2 from 11824 to 11855 genes, C3 from 11850 to 11872 genes and C4 from 12783 to 12796 genes.

D. VALIDATION AND INTERPRETATION

Using a microarray-based approach, Hou *et al.* [14] analyzed the cellular gene expression profile of lung adenocarcinoma A549 cells upon treatment with non-thermal

TABLE 6. Matrix of gene change proposed by Hou *et al.* [14].

Condition to Condition	*SE 4hr*	*LE 1hr*	*LE 2hr*	*LE 4hr*
Differentially expressed genes	*802*	*10*	*132*	*773*

TABLE 7. Matrix of gene change proposed by our approach.

Condition to Condition	*NT - SE 4hr*	*NT - LE 1hr*	*LE 1hr - LE 2hr*	*LE 2hr - LE 4hr*
Unstable Genes	*37162*	*36999*	*4039*	*4938*

plasma. They focused on finding plasma-associated molecular signatures to elucidate the impact of NTP on the transcriptome of this tumor cell.

Even though survival of the 3-min treatment group decreased to only approximately 20 % at 4h post exposure, when compared to sham control, the RNA integrity number (RIN) still showed that RNA was not degraded and had sufficiently high quality for further analysis (data not shown).

With the selection criteria mentioned above, 1209 differentially expressed genes were obtained for all time points. Specifically, at 4h after 1-min NTP treatment, 802 genes (559 down-regulated and 243 genes) showed significant expression according to the preset criteria (with fold changes more than 1.2, and FDR p-value less than 0.05), whereas only 10 genes (10 down-regulated genes), 132 genes (109 down-regulated and 23 up-regulated genes) and 773 genes (684 down-regulated and 89 up-regulated genes) expressed at 1h, 2h and 4h, respectively, after 3-min NTP exposures. These data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE59997. A complete overview of the differentially regulated genes can be obtained using GEO2R or other software. Table 6 summarizes the above-described results.

Based on our approach, the number of genes changing their cluster membership between two states or conditions is presented in Table 7. It is seminal to clarify that this gene membership comparison is based on the fusion of the three states under consideration. That is, this comparison is for the fusion of NT, fusion of SE, fusion of LE states, and not for samples 1, 2 and 3 of each condition. The comparison reveals interesting results, with 36999 genes changing cluster membership from NT to LE post 1hr. There is also an eminent decline in the number of genes that changed their cluster membership from LE post 1hr to LE post 2hr, to only 4039 genes. Further, 899 more genes joined the 4039 genes to change their cluster membership from LE post 2hr to LE post 4hr.

Non-small cell lung cancer is defined at the molecular level by mutations and alterations to oncogenes including AKT1, ALK, BRAF, EGFR, HER2, KRAS, MEK1, MET, NRAS, PIK3CA, RET and ROS1. In Table 8, we present a molecular expression profiling of the biological biomarkers for non-small cell lung cell cancer due to fusion of NT, LE post 1hr, LE post 2hr and LE post 4hr.

TABLE 8. NSLC oncogenes molecular profiling.

	Fusion of NT	*Fusion of LE post 1hr*	*Fusion of LE post 2hr*	*Fusion of LE post 4hr*
AKT1	*4*	*1*	*1*	*1*
AKT1	*1*	*1*	*1*	*1*
AKT1	*2*	*3*	*3*	*3*
MET	*1*	*3*	*3*	*3*
MET	*1*	*4*	*4*	*4*
MET	*4*	*4*	*4*	*4*
MET	*1*	*3*	*4*	*4*
KRAS	*1*	*3*	*3*	*3*
KRAS	*4*	*1*	*1*	*2*
KRAS	*1*	*2*	*2*	*2*
KRAS	*4*	*4*	*4*	*4*
NRAS	*4*	*1*	*1*	*1*
NRAS	*1*	*3*	*3*	*3*
EGFR	*1*	*3*	*3*	*3*
EGFR	*4*	*4*	*4*	*4*
EGFR	*4*	*1*	*1*	*1*
EGFR	*1*	*2*	*2*	*2*
EGFR	*2*	*3*	*3*	*3*
EGFR	*3*	*2*	*2*	*2*
EGFR	*1*	*1*	*1*	*1*
EGFR	*2*	*1*	*1*	*1*
BRAF	*1*	*2*	*2*	*2*
ALK	*2*	*4*	*4*	*4*
ALK	*2*	*4*	*4*	*2*
PIK3CA	*1*	*3*	*3*	*3*
PIK3CA	*1*	*1*	*1*	*1*
SMEK1	*4*	*2*	*2*	*2*
RET	*2*	*2*	*2*	*2*
RET	*2*	*3*	*3*	*3*
RET	*2*	*3*	*3*	*3*
ROS1	*2*	*3*	*3*	*1*

Genes BRAF, RET and ROS1 show an overexpression transition by one cluster from NT to LE states, followed by a stabilization in the 3 LE states at the same cluster. BRAF move from C1 to C2 while both RET and ROS1 move from C2 to C3.

Genes MET, ALK and PIK3CA show an overexpression transition by two clusters from NT to LE states, and a stabilization in the 3 LE states at the same cluster. MET moves from C1 to C3 and C4, ALK moves from C2 to C4 while PIK3CA moves from C1 to C3.

SME1 shows an under-expression transition by two clusters from NT to LE states, followed by a stabilization in the 3 LE states at the same cluster. SME1 moves from C4 to C2.

VI. CONCLUSION

In this paper, we presented a novel fusion-based approach for gene expression profiling in non-small cell lung cancer. The goal was to discover genes related to non-small cell lung cancer. The main contribution was to exploit Dempster-Shafer evidence theory-based data fusion to integrate data from different cases in the considered dataset to provide clinicians with more consistent and accurate information, than can be offered by any individual case. In addition, we proposed use of fuzzy c-means-with-range clustering to track changes of genes' states under different non-thermal plasma treatments.

Experiments depicted potential biological biomarkers for tumorigenesis of lung cancer. Oncogenes BRAF, RET, ROS1, MET, ALK and PIK3CA were found to exhibit an overexpression transition from non-treatment to non-thermal plasma treatment state. SME1 is the only gene exhibiting a suppression from the NT state to the LE states. However, an eminent gene expression consistency in the non-thermal plasma treatment state at post 1hr, 2hr and 4hr was observed for all genes.

For future work, we plan to use more cases studies and test the performance of our proposed approach on different datasets. Additionally, we plan to introduce a mathematical model that can simulate dynamics of gene expression profiles and integrate it with qualitative counterpart evidence fusion clustering. Further, our current work on uncertainty modeling can be extended as it is a first attempt to model uncertainty related to gene expression based on data fusion. Our proposed process is composed of many steps, with each step is a source of uncertainty. This uncertainty can be propagated from one-step to another, which can influence the final decision on genes related to non-small cell lung cancer. Integrating uncertainty propagation in our proposed approach will constitute a further challenging perspective for other related works [27]–[29].

REFERENCES

[1] A. Wang, N. An, G. Chen, L. Liu, and G. Alterovitz, "Subtype dependent biomarker identification and tumor classification from gene expression profiles," *Knowl.-Based Syst.*, vol. 146, pp. 104–117, Apr. 2018.

[2] D. Jiang, C. Tang, and A. Zhang, "Cluster analysis for gene expression data: A survey," *IEEE Trans. Knowl. Data Eng.*, vol. 16, no. 11, pp. 1370–1386, Nov. 2004.

[3] K. Cao and Q. Zhu, "Belief combination for uncertainty reduction in microarray gene expression pattern analysis," in *Proc. Int. Conf. Comput. Sci.*, 2007, pp. 844–851.

[4] P. Langfelder, B. Zhang, and S. Horvath, "Defining clusters from a hierarchical cluster tree: The dynamic tree cut package for R," *Bioinformatics*, vol. 24, no. 5, pp. 719–720, Mar. 2008.

[5] A. L. Richards, P. Holmans, M. C. O'Donovan, M. J. Owen, and L. Jones, "A comparison of four clustering methods for brain expression microarray data," *BMC Bioinf.*, vol. 9, no. 1, p. 490, 2008.

[6] K. Wang, J. Zheng, J. Zhang, and J. Dong, "Estimating the number of clusters via system evolution for cluster analysis of gene expression data," *IEEE Trans. Inf. Technol. Biomed.*, vol. 13, no. 5, pp. 848–853, Sep. 2009.

[7] H. Zhang and Y.-F. Sun, "Learning gene regulatory networks based on Dempster–Shafer evidence theory," in *Proc. 3rd Int. Conf. Adv. Comput. Theory Eng.*, 2010, pp. 100–104.

[8] M. Krejník and J. Kléma, "Empirical evidence of the applicability of functional clustering through gene expression classification," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 9, no. 3, pp. 788–798, May/Jun. 2012.

[9] X. Chen and C. Jian, "Gene expression data clustering based on graph regularized subspace segmentation," *Neurocomputing*, vol. 143, pp. 44–50, Nov. 2014.

[10] Z. Jiang, T. Li, W. Min, Z. Qi, and Y. Rao, "Fuzzy c-means clustering based on weights and gene expression programming," *Pattern Recognit. Lett.*, vol. 90, pp. 1–7, Apr. 2017.

[11] P. Dutta and S. Saha, "Fusion of expression values and protein interaction information using multi-objective optimization for improving gene clustering," *Comput. Biol. Med.*, vol. 89, pp. 31–43, Oct. 2017.

[12] A. K. Paul and P. C. Shill, "Incorporating gene ontology into fuzzy relational clustering of microarray gene expression data," *Biosystems*, vol. 163, pp. 1–10, Jan. 2018.

[13] D. Dhar. (Apr. 2009). "States of matter." [Online]. Available: https://arxiv.org/abs/0904.2664

[14] J. Hou *et al.*, "Non-thermal plasma treatment altered gene expression profiling in non-small-cell lung cancer A549 cells," *BMC Genomics*, vol. 16, no. 1, p. 435, Dec. 2015.

[15] A. Banerjee and R. N. Dave, "Validating clusters using the hopkins statistic," in *Proc. Int. Conf. Fuzzy Syst.*, vol. 1, Jul. 2004, pp. 149–153.

[16] E. Panayirci and R. C. Dubes, "A test for multidimensional clustering tendency," *Pattern Recognit.*, vol. 16, no. 4, pp. 433–444, 1983.

[17] J. C. Bezdek, R. J. Hathaway, and J. M. Huband, "Visual assessment of clustering tendency for rectangular dissimilarity matrices," *IEEE Trans. Fuzzy Syst.*, vol. 15, no. 5, pp. 890–903, Oct. 2007.

[18] J. C. Bezdek, R. Ehrlich, and W. Full, "FCM: The fuzzy c-means clustering algorithm," *Comput. Geosci.*, vol. 10, nos. 2–3, pp. 191–203, 1984.

[19] M.-S. Yang and Y. Nataliani, "Robust-learning fuzzy c-means clustering algorithm with unknown number of clusters," *Pattern Recognit.*, vol. 71, pp. 45–59, Nov. 2017.

[20] I. R. Farah, W. Boulila, K. S. Ettabaa, B. Solaiman, and M. Ben Ahmed, "Interpretation of multisensor remote sensing images: Multiapproach fusion of uncertain information," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 12, pp. 4142–4152, Dec. 2008.

[21] I. R. Farah, W. Boulila, K. S. Ettabaa, and M. Ben Ahmed, "Multiapproach system based on fusion of multispectral images for land-cover classification," *IEEE Trans. Geosci. Remote Sens.*, vol. 46, no. 12, pp. 4153–4161, Dec. 2008.

[22] R. R. Yager, "A class of fuzzy measures generated from a Dempster–Shafer belief structure," *Int. J. Intell. Syst.*, vol. 14, no. 12, pp. 1239–1247, Dec. 1999.

[23] R. R. Yager, "Satisfying uncertain targets using measure generalized Dempster–Shafer belief structures," *Knowl.-Based Syst.*, vol. 142, pp. 1–6, Feb. 2018.

[24] W. Boulila, I. R. Farah, K. S. Ettabaa, B. Solaiman, and H. Ben Ghézala, "Improving spatiotemporal change detection: A high level fusion approach for discovering uncertain knowledge from satellite image databases," in *Proc. Int. Conf. Data Mining*, vol. 9, Oct. 2009, pp. 222–227.

[25] T. Denœux, S. Li, and S. Sriboonchitta, "Evaluating and comparing soft partitions: An approach based on Dempster–Shafer theory," *IEEE Trans. Fuzzy Syst.*, vol. 26, no. 3, pp. 1231–1244, Jun. 2018.

[26] I. Bloch, "Information combination operators for data fusion: A comparative review with classification," *IEEE Trans. Syst., Man, Cybern. A, Syst. Humans*, vol. 26, no. 1, pp. 52–67, Jan. 1996.

[27] A. Ferchichi, W. Boulila, and I. R. Farah, "Towards an uncertainty reduction framework for land-cover change prediction using possibility theory," *Vietnam J. Comput. Sci.*, vol. 4, no. 3, pp. 195–209, Aug. 2017.

[28] A. Ferchichi, W. Boulila, and I. R. Farah, "Propagating aleatory and epistemic uncertainty in land cover change prediction process," *Ecol. Inform.*, vol. 37, pp. 24–37, Jan. 2017.

[29] W. Boulila, Z. Ayadi, and I. R. Farah, "Sensitivity analysis approach to model epistemic and aleatory imperfection: Application to land cover change prediction model," *J. Comput. Sci.*, vol. 23, pp. 58–70, Nov. 2017.

MUHAMED WAEL FAROUQ received the M.Sc. and Ph.D. degrees in applied statistics from the Faculty of Commerce, Ain Shams University, in 2010 and 2018, respectively, where he has been an Assistant Lecturer of statistics, since 2010. He has been a Visiting Researcher with the University of Stirling, U.K., from 2015 to 2017. His research interests include computational statistics, biostatistics, mathematical biology, population dynamics, multivariate analysis, and data mining. He has published papers in refereed journals and conference proceedings in these areas. During this period, he was jointly conducting his Ph.D. research with the Cognitive Big Data Informatics (CogBID) Lab (led by Prof. A. Hussain), and acted as a Manager of the Lab for the same period. In addition, he has been acting as the Chair of the IEEE University of Stirling Student Branch, from 2015 to 2017.

WADII BOULILA received the Engineering degree (Hons.) in computer science from the Aviation School of Borj El Amri, in 2005, the M.Sc. degree from the National School of Computer Science (ENSI), University of Manouba, Tunisia, in 2007, and the Ph.D. degree conjointly from ENSI and Telecom Bretagne, University of Rennes 1, France, in 2012. From 2012 to 2015, he was an Assistant Professor in computer science with the Higher Institute of Multimedia Arts of Manouba, Manouba University, Tunisia. He is currently an Assistant Professor of computer science with the IS Department, College of Computer Science and Engineering, Taibah University, Saudi Arabia. He is also a Permanent Researcher with the RIADI Laboratory, University of Manouba, and an Associate Researcher with the ITI Department, University of Rennes 1, France. His primary research interests include big data analytics, deep learning, data mining, artificial intelligence, uncertainty modeling, and remote sensing images. He has served as the Chair, Reviewer, and a TPC Member for many leading international conferences and journals.

MEDHAT ABDEL-AAL was the Former Head of the Statistics, Mathematics and Insurance Department, Faculty of Commerce, from 2014 to 2015. He is currently a Professor of statistics with the Faculty of Commerce, Ain Shams University, Egypt. His research interests include data mining and knowledge discovery, machine learning, and pattern classification. He has published papers in refereed journals and conference proceedings in these areas. He has authored and co-authored of many books in English and Arabic Languages.

AMIR HUSSAIN (SM'97) received the B.Eng. (Hons.) and Ph.D. degrees in electronic and electrical engineering from the University of Strathclyde, Glasgow, U.K., in 1992 and 1997, respectively.

Following the postdoctoral and academic positions held at the Universities of West of Scotland (1996–1998), Dundee (1998–2000), and Stirling (2000–2018), respectively, he joined Edinburgh Napier University, U.K., in 2018, as the founding Director of the Cognitive Big Data and Cybersecurity (CogBiD) Research Lab. His research interests are cross-disciplinary and industry-focused, aimed at pioneering brain-inspired, cognitive Big Data technology for solving complex real-world problems. He has co-authored nearly 400 publications, with around 150 journal papers and over a dozen Books. He has led major multi-disciplinary research projects, as a Principal Investigator, funded by the national and European research councils, local and international charities, and industry. Until now, he has supervised more than 30 PhDs. He is a Senior Fellow of the Brain Sciences Foundation, Wellesley, MA, USA. In 2017, he was ranked, in an independent survey (published in the *Information Processing and Management Journal* (Elsevier)), as one of the world's top two most productive researchers in sentiment analytics, since 2000. Amongst other distinguished roles, he is the General Chair of the IEEE WCCI 2020 (the world's largest and top IEEE Technical Event in Computational Intelligence, comprising IJCNN, FUZZ-IEEE, and the IEEE CE), and the Vice-Chair of the Emergent Technologies Technical Committee of the IEEE Computational Intelligence Society. He is the Chapter Chair of the IEEE UK & RI Industry Applications Society Chapter. He is the founding Editor-in-Chief of the *Cognitive Computation journal* (Springer Nature) and the *BMC Big Data Analytics journal* (BioMed Central (BMC) – part of Springer Nature). He has been appointed as an Associate Editor of several other world-leading journals including, the IEEE TRANSACTIONS ON NEURAL NETWORKS AND LEARNING SYSTEMS, the *Information Fusion journal* (Elsevier), the IEEE TRANSACTIONS ON EMERGING TOPICS IN COMPUTATIONAL INTELLIGENCE, and the *IEEE Computational Intelligence Magazine*.

ABDEL-BADEEH SALEM was a Director of the Scientific Computing Center, Ain Shams University, Cairo, Egypt, from 1984 to 1990, where he was a Professor of computer science with the Faculty of Science, from 1989 to 1996, a former Vice Dean of the Faculty of Computer and Information Sciences, from 1996 to 2007, and has been a Professor Emeritus of Computer Science, since 2007. His research interests include intelligent computing, expert systems, medical informatics, and intelligent e-learning technologies. He has published around 200 papers in refereed journals and conference proceedings in these areas. He has been involved in more than 200 conferences and workshops as an Int. Program Committee, Organizer, and Session Chair. He has authored and co-authored of 15 books in English and Arabic Languages. He also served as a Member of many scientific societies. He is a member of the Editorial Board of many leading international conferences and journals.