IEEE Access
Multidisciplinary : Rapid Review : Open Access Journal

Received February 4, 2019, accepted March 30, 2019, date of publication April 11, 2019, date of current version April 25, 2019.

Digital Object Identifier 10.1109/ACCESS.2019.2910078

Overlapping Clusters and Support Vector Machines Based Interval Type-2 Fuzzy System for the Prediction of Peptide Binding Affinity

VOLKAN USLAN[1], **(Member, IEEE), HUSEYIN SEKER**[2], **(Member, IEEE), AND ROBERT JOHN**[3], **(Senior Member, IEEE)**

[1]School of Computer Science and Informatics, De Montfort University, Leicester LE1 9BH, U.K.
[2]Department of Computer Science and Digital Technologies, University of Northumbria at Newcastle, Newcastle upon Tyne NE1 8ST, U.K.
[3]School of Computer Science, University of Nottingham, Nottingham NG8 1BB, U.K.

Corresponding author: Huseyin Seker (huseyin.seker@northumbria.ac.uk)

This work was supported by De Montfort University (DMU), Leicester, U.K.

ABSTRACT In the post-genome era, it is becoming more complex to process high dimensional, low-instance available, and nonlinear biological datasets. This paper aims to address these characteristics as they have adverse effects on the performance of predictive models in bioinformatics. In this paper, an interval type-2 Takagi Sugeno fuzzy predictive model is proposed in order to manage high-dimensionality and nonlinearity of such datasets which is the common feature in bioinformatics. A new clustering framework is proposed for this purpose to simplify antecedent operations for an interval type-2 fuzzy system. This new clustering framework is based on overlapping regions between the clusters. The cluster analysis of partitions and statistical information derived from them has identified the upper and lower membership functions forming the premise part. This is further enhanced by adapting the regression version of support vector machines in the consequent part. The proposed method is used in experiments to quantitatively predict affinities of peptide bindings to biomolecules. This case study imposes a challenge in post-genome studies and remains an open problem due to the complexity of the biological system, diversity of peptides, and curse of dimensionality of amino acid index representation characterizing the peptides. Utilizing four different peptide binding affinity datasets, the proposed method resulted in better generalization ability for all of them yielding an improved prediction accuracy of up to 58.2% on unseen peptides in comparison with the predictive methods presented in the literature. Source code of the algorithm is available at https://github.com/sekerbigdatalab.

INDEX TERMS Interval type-2 fuzzy systems, support vector regression, overlapping clusters, peptide binding affinity, clustering, high-dimensionality.

I. INTRODUCTION

Peptides, a small sequence of amino acids, often interacts with proteins in cellular processes [1]. One of the important peptide-protein interactions occur when a peptide binds to a Major Histocompatibility Complex (MHC) forming a peptide-MHC (pMHC) complex. pMHC is transported to the cell membrane where it is recognized by a T-cell in order to induce an immune response. Therefore, in pharmaceutical studies, validation of a pMHC binding with the drug of interest is crucial. However, this is a complicated process and computational methods are constantly being developed to support traditional empirical research to identify most likely candidates out of a library of thousands of peptides. Moreover, predictive models based on sequence-based methods are needed to approximate the binding affinities.

In recent years, the problem of binding affinity predictions became two-fold. Qualitative studies consider classifying binding predictions as 'binders' and 'non-binders' [2] or 'weak' and 'strong' binders [3]–[5] whereas quantitative studies allow real-value binding predictions [6]. Lately, regression-based approaches have become more prevalent in sequence-based studies. A number of methods are used as predictors such as the partial least squares [7], random forests [8], support vector regression [9] and regularization methods [10]. Nevertheless, the complexity of a biological

The associate editor coordinating the review of this manuscript and approving it for publication was Chee Keong Kwoh.

system, diversity of peptides, and curse of dimensionality of amino acid index representation that characterise the peptides have adverse effects on the performance of peptide-binding predictive models. Moreover, uncertainties are prevalent in peptide binding affinity datasets due to imprecise or noisy measurements, and these datasets need to be analysed appropriately [11]. There is still a lack of methods accounting for this aspect of peptide-protein bindings [12].

In certain applications, where the data is complex and non-linear, fuzzy systems are more tolerant of imprecise information and capable of modelling linguistic and numerical uncertainty. Moreover, they form a rule-based structure similar to human reasoning. Presently, type-2 fuzzy systems [13] have a wider use in real-world applications than ever before [14]. They, in certain applications, perform better than type-1 fuzzy systems in terms of modelling and minimizing uncertainties [15]–[17]. Type-2 fuzzy systems are preferred due to the consideration of membership functions being imprecise and being able to cope with the uncertainties associated with them.

In this paper, an overlapping clusters and support vector machine based interval type-2 Takagi Sugeno fuzzy system is proposed to address the aforementioned shortcomings of the sequence-based predictive models. A novel clustering framework is proposed in order to simplify antecedent operations for an interval type-2 fuzzy system. This clustering framework is based on overlapping regions between the clusters. The cluster analysis of partitions and statistical information derived from them have identified the upper and lower membership functions forming the premise part. This is further enhanced by adapting the regression version of support vector machines (SVR) in the consequent part [18]. The computational demand in the defuzzification process is addressed by a method which has the closed-form representation. In addition, feature selection is used in order to reduce the high number of amino acid biochemical descriptors, representing a peptide, which formed the input scheme of the learning model. The prediction results indicate that the proposed model not only minimized the effects of uncertain continuous peptide binding affinities but also provided high precision in unravelling the binding affinities of unobserved peptides.

The remainder of the study starts with introducing the materials and methods (Section II). This section describes the identification of SVR based interval type-2 fuzzy system with overlapping clusters concept. Section III shows the results of the case study along with the discussion. Finally, concluding remarks are given in Section IV.

II. MATERIALS AND METHODS

A. SUPPORT VECTOR-BASED INTERVAL TYPE-2 FUZZY SYSTEM

Type-2 fuzzy sets, which are defined through membership functions, are themselves fuzzy. However, the computations of type-2 fuzzy sets are complex and in order to ease these computations Interval Type-2 (IT2) fuzzy sets can be used [19]. The Takagi Sugeno model is one of the widely used fuzzy systems [20]. This model structure presents the design of consequent parameters using a linear function. The rule-base of the interval type-2 Takagi Sugeno fuzzy system with r rules can be expressed as:

$$\begin{aligned} &R^i : \text{IF } x_1 \text{ is } \tilde{A}_1^i \text{ and } x_2 \text{ is } \tilde{A}_2^i \ldots \text{and } x_n \text{ is } \tilde{A}_n^i \\ &\text{THEN } y_i = c_{0i} + c_{1i}x_1 + \ldots + c_{ni}x_n \end{aligned} \tag{1}$$

where, x_1, x_2,..., x_n represent the input vector and c_0, c_1, c_2,..., c_n are the regression coefficients; IT2 fuzzy set is denoted by $\tilde{A}_n^i$ for the variable n and rule i; and y_i is the rule output.

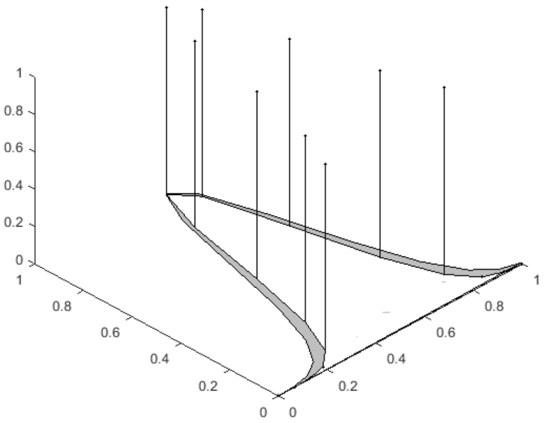


FIGURE 1. An interval type-2 fuzzy set.

Type-2 fuzzy sets should be placed in the premise or consequent part (or both) in order to define a type-2 fuzzy system. IT2 fuzzy sets are characterized by the upper membership functions (UMFs) and lower membership functions (LMFs). This is how the uncertainty is modeled for the IT2 membership function. Bounded region between UMF and LMF is the footprint of uncertainty (FOU). Each interval type-2 fuzzy set within the footprint of uncertainty is unity. Three-dimensional representation of an interval type-2 fuzzy set is depicted in Fig. 1. The firing strengths of interval type-2 fuzzy system are determined by using the t-norm operator and can be calculated as:

$$\underline{f_i} = \prod_{k=1}^{n} \underline{\mu}(x_k) \tag{2}$$

$$\overline{f_i} = \prod_{k=1}^{n} \overline{\mu}(x_k) \tag{3}$$

where $\underline{f_i}$ ($\overline{f_i}$) is the lower (upper) firing strength; $\underline{\mu}(x_k)$ ($\overline{\mu}(x_k)$) is the lower (upper) membership degree for input variable x_k; respectively, and $\prod$ denotes the product t-norm operation.

The output of an IT2 fuzzy system is obtained through type-reduction and defuzzification. The Karnik-Mendel algorithm is the widely used type-reduction method that can compute the left and right end points required for the

IT2 fuzzy set [21]. Then these end points are defuzzified to get the final output. Karnik-Mendel is an iterative algorithm and suffers from time intense computations. Therefore, alternate approaches have been presented in the literature [22]–[24]. However, the proposed IT2 fuzzy system implements Biglarbegian-Melek-Mendel (BMM) method [25] which has the closed mathematical form as described in (4)

$$Y_{\mathrm{BMM}} = q\frac{\sum_{i=1}^{r}\overline{f_i}\, y_i}{\sum_{i=1}^{r}\overline{f_i}} + p\frac{\sum_{i=1}^{r}\underline{f_i}\, y_i}{\sum_{i=1}^{r}\underline{f_i}} \tag{4}$$

where q and p are the parameters used to design the upper and lower weighted average of the rule consequents, respectively.

Recently, support vector machines are incorporated with interval type-2 fuzzy systems to identify the parameters of the consequent part [26], [27]. The regression coefficients ($\vec{w}$ and b) that weighs the linear SVR are obtained by the training samples. To incorporate SVR with the interval type-2 fuzzy system, the input for each data item as in (5) is transformed to (6). The coefficients of rule consequents ($\vec{w}$) and b are computed using the linear SVR. For this purpose, a library for support vector machines was used [28]. Then, the output of support vector-based interval type-2 fuzzy system (y'') is obtained from (7) and (8).

$$\vec{x} = [x_1, \ \ldots, x_n] \tag{5}$$

$$\vec{x}'' = [q\overline{f}_1 + p\underline{f}_1, \ q\overline{f}_1 x_{11} + p\underline{f}_1 x_{11}, \ \ldots, \ q\overline{f}_r x_{rn} + p\underline{f}_r x_{rn}] \tag{6}$$

$$y_i'' = w_0 + \sum_{k=1}^{n}(w_i x_i) \tag{7}$$

$$y'' = q\frac{\sum_{i=1}^{r}\overline{f_i}\, y_i''}{\sum_{i=1}^{r}\overline{f_i}} + p\frac{\sum_{i=1}^{r}\underline{f_i}\, y_i''}{\sum_{i=1}^{r}\underline{f_i}} + b \tag{8}$$

B. IDENTIFICATION OF INTERVAL TYPE-2 FUZZY SETS WITH OVERLAPPING CLUSTERING CONCEPT

This section will introduce a novel method based on the overlapping clusters concept in order to initialise the interval type-2 membership function parameters. The FOU of an interval type-2 fuzzy set can be defined by varying either the mean (see Fig. 3) or the standard deviation (see Fig. 4) of the Gaussian membership function. As the overlapping regions between the clusters applicable to the latter approach, the footprint of uncertainty is formed with fixed mean and blurred standard deviations. Once the interval $[\sigma_1, \sigma_2]$ is determined, upper and lower Gaussian membership functions are obtained as follows:

$$\overline{\mu}(x) = \exp\left[-\frac{(x-c)^2}{2(\sigma_2)^2}\right] \tag{9}$$

$$\underline{\mu}(x) = \exp\left[-\frac{(x-c)^2}{2(\sigma_1)^2}\right] \tag{10}$$

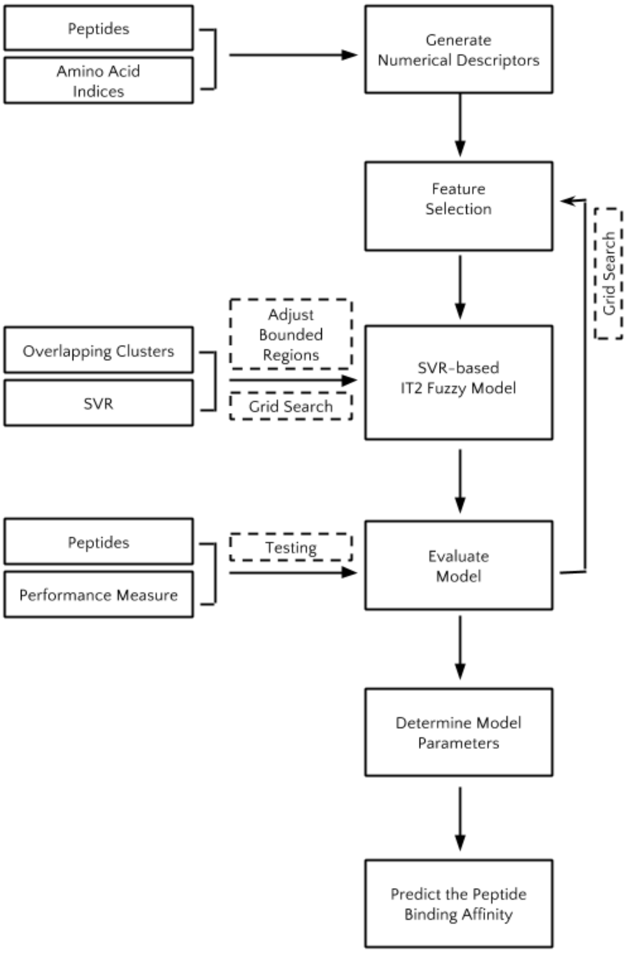


FIGURE 2. Stages of the proposed interval type-2 fuzzy system for the prediction of peptide binding affinity.

The issues that need to be considered during the system identification for a fuzzy system using clustering can be found in [29]. We considered finding interval type-2 membership function parameters with clustering methods such as the soft clustering (e.g., fuzzy c-means clustering [30]) and the crisp clustering methods (e.g., hard c-means clustering [31], hierarchical cluster analysis [32]). Statistical characteristics of clusters are used to identify the membership functions. It is assumed that statistical information that characterises a crisp cluster will involve more knowledge to identify an interval type-2 membership function than the arbitrary initialisation.

After the cluster analysis was performed we used left, right end points and centre of each cluster to define its triangular membership function. Algorithm 1 outlines the steps for finding the end points and the centre of upper and lower membership functions using the overlapping clusters concept. The proposed overlapping clusters method derives the lower membership function from the provided upper membership function approach [33], [34]. Fig. 5 illustrates how the interval type-2 fuzzy sets are formed based on the overlapping clusters as a single input-single output scheme.

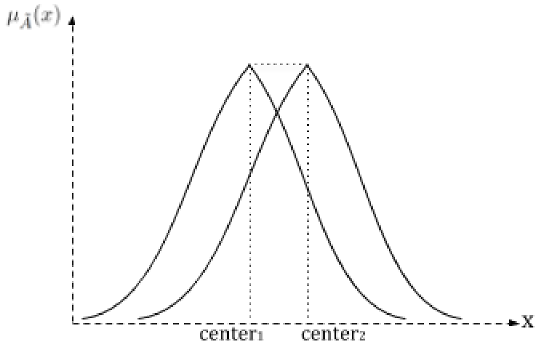


FIGURE 3. Footprint of uncertainty of an interval type-2 fuzzy set when the standard deviation is fixed and the center is blurred.

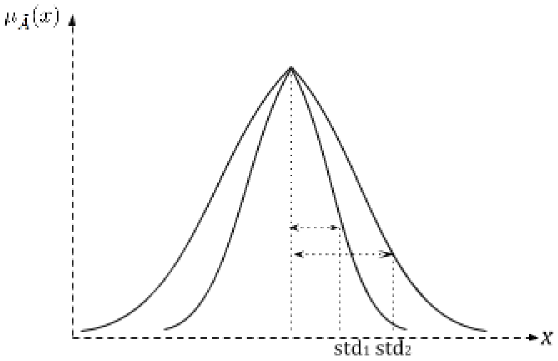


FIGURE 4. Footprint of uncertainty of an interval type-2 fuzzy set when the center is fixed and the standard deviation (std) is blurred.

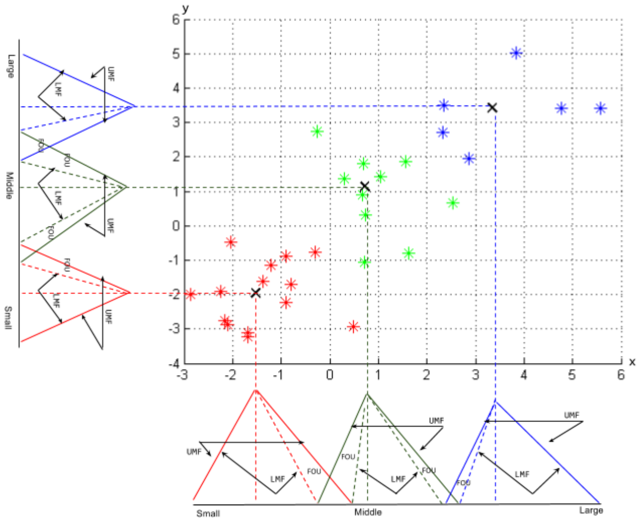


FIGURE 5. Illustration of overlapping clustering concept used to identify the end points of the interval type-2 membership function. FOU: Footprint of Uncertainty. UMF: Upper Membership Function. LMF: Lower Membership Function.

Overlaps between clusters are projected into one-dimensional data points. Neighbour clusters located on either side of the cluster identify the lower left and right end points. As a result, these end points along with the cluster centre define the parameters of the lower membership function.

Algorithm 1 Finding the End Points of the Overlapping Clusters.

```
Output: upper and lower end points of the overlapping
        clusters
set the number of clusters;
apply the clustering method;
obtain statistics of each cluster;
foreach cluster do
    set upper left point = min(cluster);
    set upper center = mean(cluster);
    set upper right point = max(cluster);
    initialize lower left point;
    set lower center = mean(cluster);
    initialize lower right point;
    foreach neighbour cluster do
        set condition left = evaluate statistical values
        (min, mean and max) of the neighbour cluster to
        find whether any of them is in the upper interval
        [left point, mean];
        if condition left then
            set lower left point = select statistics value
            found which is closer to the upper left point;
        end
        set condition right = evaluate statistical values
        (min, mean and max) of the neighbour cluster to
        find whether any of them is in the upper interval
        [mean, right point];
        if condition right then
            set lower right point = select statistics value
            found which is closer to the upper right
            point;
        end
    end
end
```

Then, we converted each triangular membership function (centre, left point, right point) to a Gaussian membership function (centre, standard deviation). The corresponding membership function may not be uniform on each wing. For a non-uniform case, even though mean remains the same for a Gaussian membership function, two separate standard deviations are required; one representing the left wing, and the other representing the right wing.

C. PEPTIDE BINDING AFFINITY DATASETS

A peptide consists of an amino acid sequence with a size of approximately 10 residues long [35]. Peptide fragments form binding with MHC class proteins as a cellular event. pMHC complexes are translocated to the membrane of the host cell where they meet T-cells. When receptors of the T-cell recognize pMHC complexes, they elicit an immune activity to happen. These immune activities range from cytotoxic killing to phagocytosis of the infected cell. One main difficulty for experimental peptide studies is that the amount of possible peptides that can bind for a particular MHC class molecule is extraordinarily large (≥ 500 billion) [36]. However, understanding how peptide-MHC class molecule interactions work and finding their binding affinities are crucial for health studies.

The proposed approach has been tested using the peptide datasets that have been obtained from various papers [37]–[40]. Each peptide dataset has been considered as a task and organized in training and test datasets [10]. For Tasks III and IV, two separate testing datasets were used even though training dataset remained the same. Table 1 lists the characteristics of the peptide binding affinity tasks. Tasks I, III and IV consist of nona-peptides whereas Task II consists of octa-peptides. Table 2 depicts the statistics of the peptide binding affinity tasks. Sequence logo (position specific amino acid frequency) representations of peptide datasets are shown in Fig. 6.

TABLE 1. Characteristics of the peptide binding affinity tasks.

Tasks	#Trng	#Test	#Total Descriptors
I	89	88	5787
II	76	76	5144
III	133	133	5787
IV	133	47	5787

#Trng: Number of peptides in the training set. #Test: Number of peptides in the testing set. #Total Descriptors: Number of total descriptors when the peptide is encoded.

TABLE 2. Statistics of the peptide binding affinity tasks.

		Tasks			
		I	II	III	IV
Min	Trng	2.94	5.01	4.30	4.30
	Test	3.13	5.01	5.08	13.00
Max	Trng	8.65	8.34	8.77	8.77
	Test	8.17	8.40	8.96	121.00
Mean	Trng	5.41	7.55	7.08	7.08
	Test	5.41	7.58	7.10	60.96
Std	Trng	1.01	0.77	0.82	0.82
	Test	0.95	0.74	0.80	33.94

Trng: Training set. Test: Testing set. Std: Standard deviation.

Amino acid feature databases such as the AAindex [41] and CISAPS [42], contain many physico-chemical and bio-chemical attributes of amino acids. Each descriptor in the amino acid feature database has twenty different numerical values along with their descriptions that correspond to each amino acid. However, previous studies usually use 643 descriptors which are mostly selected from the amino acid feature databases. To be consistent, we have encoded each amino acid in a peptide with 643 descriptors as shown in Fig. 7. The number of total descriptors becomes 5144 (643×8) and 5787 (643×9) when octa-peptide sequence and nona-peptide sequence are encoded, respectively.

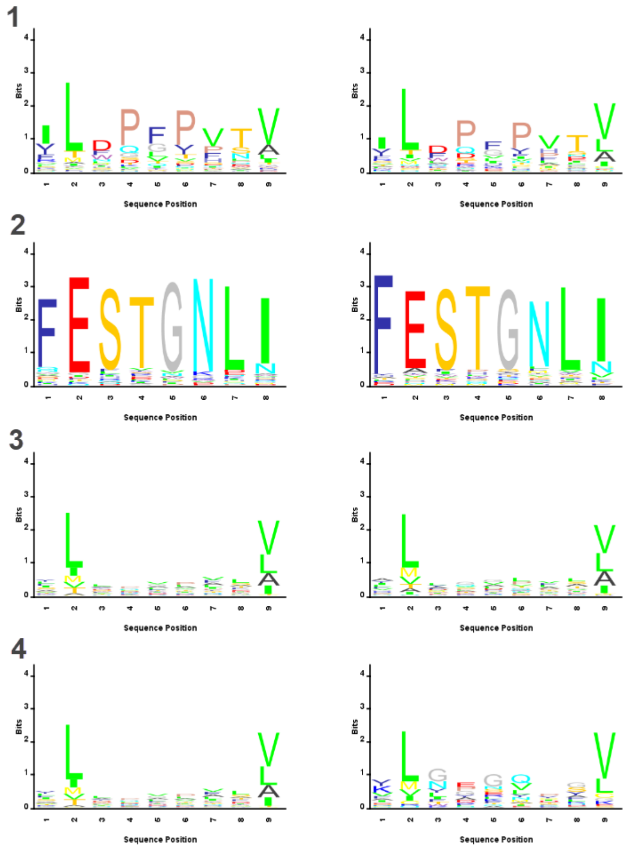


FIGURE 6. Sequence logo plots of Tasks 1-4. Training (left) and test (right) peptide datasets are represented in position specific amino acid frequencies.

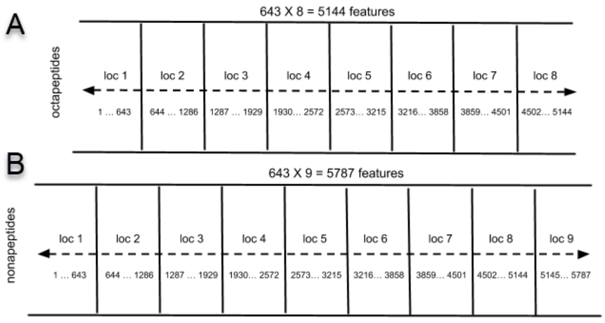


FIGURE 7. Encoding of a peptide sequence as amino acid descriptors. A) octa-peptide amino acid composition B) nona-peptide amino acid composition.

III. RESULTS AND DISCUSSION

This section presents the experimental results of overlapping clusters and support vector based interval type-2 fuzzy system that conducted on peptide binding datasets to predict the

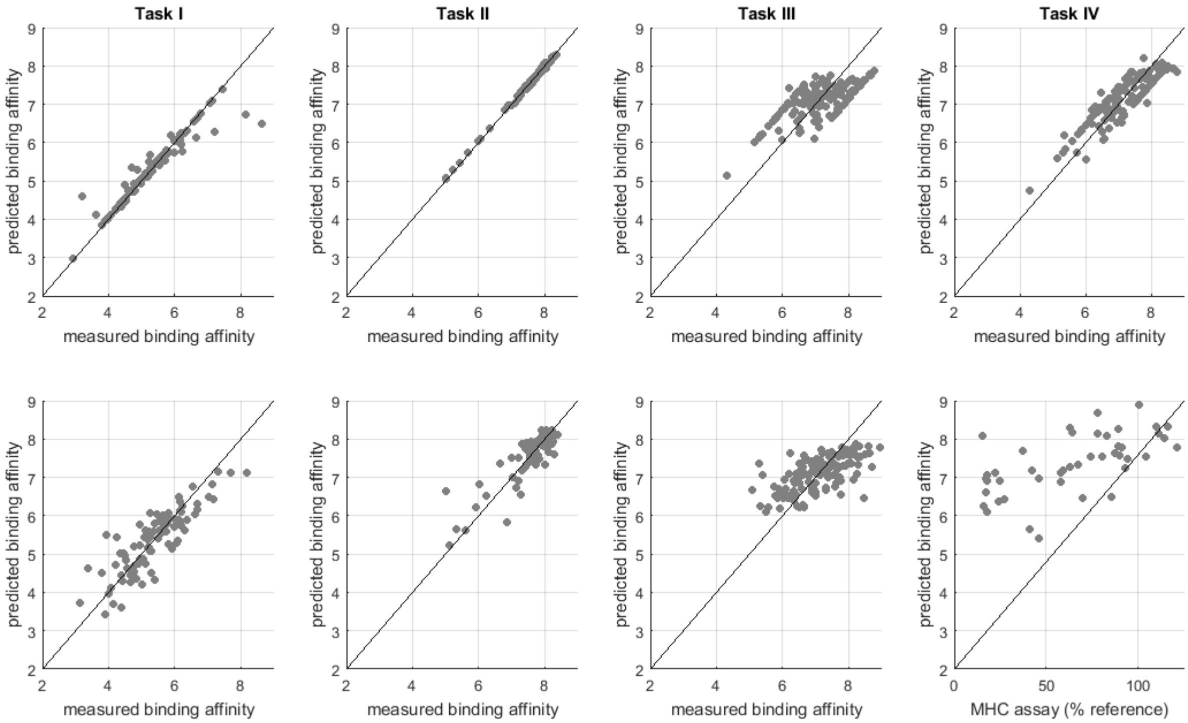


FIGURE 8. The correlation between measured and predicted peptide binding affinities; the training set is the former and the testing set is the latter.

real value of affinities. The stages of the proposed interval type-2 fuzzy model are illustrated in Fig. 2. In our implemented fuzzy model structure, type-2 fuzzy sets are in the premise and rule consequents are crisp numbers. Interval type-2 fuzzy sets of the proposed approach are determined using the overlapping clusters concept. During the system identification process of the fuzzy rule base, membership function parameter values are characterized using different clustering methods. The statistics found at the end of the cluster analysis generated the upper and lower membership functions of the interval type-2 fuzzy model. Additionally, support vector regression is used to learn the parameters of rule consequents. SVR not only enhanced the learning capability of the proposed model but also decreased the effects of overfitting. For the defuzzification process, Biglarbegian-Melek-Mendel method, which has the closed-form representation was used. We used grid search in order to find the SVR and Biglarbegian-Melek-Mendel method design parameters for the proposed interval type-2 fuzzy system.

Blind validation experiments were implemented to reveal the accuracy performance of the proposed method. Each peptide in both training and testing peptide datasets are encoded into physico-chemical and bio-chemical descriptor vectors. Then, the descriptors were normalized using min-max scaling so that every descriptor varied in the range between 0 and 1. When there is a large number of features available, feature selection is often required in bioinformatics to get rid of irrelevant features, avoid overfitting and provide an improvement in model performance [43]. As the encoded feature set became large (≥ 5000), a feature selection method (multi-cluster feature selection) is considered to be used in this work [44]. Multi-cluster feature selection is an unsupervised feature selection method that does not require labeled data and already used in many bioinformatics applications [45]–[47]. We decreased the high-dimensionality from many thousands to a few hundred. We found 161, 247, 172 and 141 descriptors are adequate to preserve a model for Tasks I, II, III and IV, respectively. It is also found that amino acid polarity appeared in the selected features of Tasks I, II and III as being the most discriminative descriptor.

To be consistent in comparisons with similar prediction methods, the coefficient of determination (q^2) [48] and the Spearman rank correlation coefficient (ρ) [49] were used. Percentage improvement of the proposed model as compared to the models found in this research domain ($I^1\%$) and to our previous work ($I^2\%$) were computed as in (11).

$$\mathrm{I}\% = |\frac{\mathrm{Model}_1 - \mathrm{Model}_2}{\mathrm{Model}_2}| \times 100 \tag{11}$$

Table 3 reports the training and testing prediction performances of the proposed method when hard c-means clustering (HCM), fuzzy c-means clustering (FCM), and hierarchical cluster analysis (HCA) were used to initialize the membership grades of the interval type-2 fuzzy sets. For all tested models, the number of clusters varied in the range between two and four. The best predictive accuracy performances are achieved with FCM (three tasks) and

TABLE 3. The prediction scores of the proposed method based on different clustering methods.

	Task I			Task II			Task III			Task IV		
Clustering	$\#k^*$	$q^2_{(trainset)}$	$q^2_{(testset)}$	$\#k^*$	$q^2_{(trainset)}$	$q^2_{(testset)}$	$\#k^*$	$q^2_{(trainset)}$	$q^2_{(testset)}$	$\#k^*$	$\rho_{(trainset)}$	$\rho_{(testset)}$
HCM	3	0.849	0.714	3	0.987	0.743	3	0.466	0.350	2	0.853	0.646
FCM	3	0.868	**0.719**	2	0.996	0.729	3	0.450	**0.367**	3	0.852	**0.659**
		SVR (C = 0.75; ϵ = 0.05)						SVR (C = 1.25; ϵ = 0.85)			SVR (C = 1.75; ϵ = 0.45)	
		BMM (q = 0.55; p = 0.50)						BMM (q = 0.45; p = 0.50)			BMM (q = 0.70; p = 0.55)	
HCA	2	0.833	0.679	4	0.996	**0.756**	2	0.473	0.278	2	0.829	0.652
					SVR (C = 1.40; ϵ = 0.05)							
					BMM (q = 0.65; p = 0.45)							

$\#k^*$: Number of Clusters.
HCM: Hard C-Means Clustering; FCM: Fuzzy C-Means Clustering; HCA: Hierarchical Cluster Analysis.
The best results are highlighted in bold font type along with their design parameters set underneath.

TABLE 4. Comparison of the results of the proposed method with respect to those reported in this research domain.

		Task I		Task II		Task III		Task IV	
Models		Number of Features	q^2	Number of Features	q^2	Number of Features	q^2	Number of Features	ρ
Lasso	[10]	50	0.667	43	0.642	56	0.205	41	0.548
Ridge Regression	[10]	50	0.691	43	0.668	56	0.131	41	0.586
Partial Least Squares	[7]	584	0.455	147	0.401	180	0.153		n/a
G-Kernel Partial Least Squares	[7]	584	0.678	147	0.746	180	0.200		n/a
E-Kernel Partial Least Squares	[7]	584	0.691	147	0.590	180	0.219		n/a
Support Vector Regression	[8]	200	0.682	100	0.639	100	0.232		n/a
Random Forests	[8]	200	0.661	200	0.607	100	0.208		n/a
SV-based T1 Fuzzy System	[12]	161	0.696	247	0.743	172	0.310	141	0.643
OCSV-based IT2 Fuzzy System		161	**0.719**	247	**0.756**	172	**0.367**	141	**0.659**
I^1%			4.1%		1.3%		58.2%		12.5%
I^2%			3.3%		1.8%		18.4%		2.5%

I^1%: Percentage improvement of the proposed model as compared to the models found in this research domain.
I^2%: Percentage improvement of the proposed model as compared to the models of our previous work.
n/a: not available.
G-Kernel: Gaussian Kernel. E-Kernel: Exponential Kernel.
SV-based T1 Fuzzy System: Support Vector-based Type-1 Fuzzy System.
OCSV-based IT2 Fuzzy System: Overlapping Clustering and Support Vector based Interval Type-2 Fuzzy System.
The best results are highlighted in bold font type.

HCA (one task). As can be seen underneath the best models, their SVR (C and ϵ) and Biglarbegian-Melek-Mendel method design parameters (q and p) were given. For all tasks, we trained SVR with a linear kernel to obtain the rule consequent coefficients of the proposed interval type-2 fuzzy system.

The correlation between measured and predicted real value binding affinities are shown in Fig. 8. The best models of the proposed method (overlapping clustering and support vector based interval type-2 fuzzy system) achieved higher accuracy and significant increase in prediction performance than the previously published methods [7], [8], [10] on unseen peptides as shown in Table 4. As compared to the best predictive methods (0.691, 0.746, 0.232 and 0.586) presented in the literature, the proposed method resulted in better generalization ability for all of them yielding an improved prediction accuracy of 4.1%, 1.3%, 58.2% and 12.5% for Tasks I, II, III and IV, respectively. Additionally, as compared to our previous work (support vector based type-1 fuzzy system) [12], the proposed method achieved an accuracy improvement in prediction performance of 3.3%, 1.8%, 18.4% and 2.5% for Tasks I, II, III and IV, respectively.

Defining fuzzy sets and the number of rules are the main concerns in structure identification of a fuzzy system. The formation of rules can be automated with the help of the cluster analysis where each partition maps to a fuzzy rule. In clustering, the parameter to indicate the number of clusters should be preset before the cluster analysis is performed. However, determining the exact number of clusters is a considerable difficulty. We performed a grid search to observe (from two to up to seven clusters) to see the tendency of groupings within the peptide binding affinity datasets. We found that mostly three clusters are the natural number of the grouping of peptide binding affinities when incorporated with the proposed interval type-2 fuzzy system. The number of clusters we found for the peptide binding affinity datasets also agree with the fact that the number of membership functions should be ≤ 7 in each input domain for the practical design of an interval type-2 fuzzy logic system [50]. This magical number is based on a study [51] stating that keeping in mind more than 7 ± 2 objects at the same time becomes more confusing for a human and beyond his/her processing information capacity.

The utilization of overlapping clusters aimed for overcoming the difficulties of parameter identification process in an interval type-2 fuzzy system. When required, interval type-2 membership function parameters can be further optimised using a learning algorithm [52]. As the initialisation of membership functions depend also the parameter values of learning algorithms, the proposed initialisation process will eliminate this necessity and lead a learning algorithm to focus its ultimate purpose.

Finally, in this study we used Gaussian membership functions as they are relatively easy to implement and require less parameters, therefore have less assumptions. However, any type of membership functions could have been used and these will be implemented in future work and tested to see if they offer any improvements over our current method.

IV. CONCLUSION

This paper presents a robust hybrid system that incorporates an overlapping clustering concept and support vector regression for the design of an interval-type-2 fuzzy system. This is one of the first studies where a support-vector based interval type-2 fuzzy system is applied to a real bioinformatics problem. The performance and robustness of the proposed hybrid predictive models were demonstrated over one of the most challenging problems in molecular biology - the prediction of peptide binding affinity. The analyses on four different case studies in the prediction of peptide binding affinity have yielded better generalisation ability and higher predictive accuracy than those presented in the literature. This study has both biological and computational implications: the predictive model has yielded a number of useful biological characteristics of the peptides (e.g. amino acid polarity) which could help analysis of peptides with more appropriate binding affinities. In addition to the development of a robust predictive model with applications in high dimensional datasets (rare in fuzzy system-based studies), the study presents a successful implementation of the overlapping clustering framework in the design of an interval type-2 fuzzy system. As this framework can also help determine initial values of the interval type-2 fuzzy system, it could be further incorporated with any type of clustering, machine learning and optimisation methods to help further improve its outcome. Further research will be carried out towards this direction.

REFERENCES

[1] H. J. Dyson and P. E. Wright, "Intrinsically unstructured proteins and their functions," *Nature Rev. Mol. Cell Biol.*, vol. 6, pp. 197–208, Mar. 2005.

[2] A. Sette *et al.*, "Prediction of major histocompatibility complex binding regions of protein antigens by sequence pattern analysis," *Proc. Nat. Acad. Sci. USA*, vol. 86, no. 9, pp. 3296–3300, 1989.

[3] J. D. Stone, A. S. Chervin, and D. M. Kranz, "T-cell receptor binding affinities and kinetics: Impact on T-cell activity and specificity," *Immunology*, vol. 126, no. 2, pp. 165–176, 2009.

[4] K. Roomp, I. Antes, and T. Lengauer, "Predicting MHC class I epitopes in large datasets," *Bioinformatics*, vol. 11, no. 1, p. 90, 2010.

[5] R. D. Bremel and E. J. Homan, "An integrated approach to epitope analysis I: Dimensional reduction, visualization and prediction of MHC binding using amino acid principal components and regression approaches," *Immunome Res.*, vol. 6, no. 1, p. 7, 2010.

[6] I. A. Doytchinova, M. J. Blythe, and D. R. Flower, "Additive method for the prediction of protein-peptide binding affinity. Application to the MHC class I molecule HLA-A0201," *J. Proteome Res.*, vol. 1, no. 3, pp. 263–272, 2002.

[7] C. Bergeron *et al.* (2011). "Prediction of peptide bonding affinity: Kernel methods for nonlinear modeling." [Online]. Available: https://arxiv.org/abs/1108.5397

[8] A. Srivastava, S. Ghosh, N. Anantharaman, and V. K. Jayaraman, "Hybrid biogeography based simultaneous feature selection and MHC class I peptide binding prediction using support vector machines and random forests," *J. Immunol. Methods*, vol. 387, nos. 1–2, pp. 284–292, 2013.

[9] W. Liu, X. Meng, Q. Xu, D. R. Flower, and T. Li, "Quantitative prediction of mouse class I MHC peptide binding affinity using support vector machine regression (SVR) models," *Bioinformatics*, vol. 7, no. 1, p. 182, 2006.

[10] O. Demir-Kavuk, M. Kamada, T. Akutsu, and E.-W. Knapp, "Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features," *Bioinformatics*, vol. 12, no. 1, p. 412, 2011.

[11] D. R. Flower, "Towards in silico prediction of immunogenic epitopes," *Trends Immunol.*, vol. 24, no. 12, pp. 667–674, 2003.

[12] V. Uslan and H. Seker, "Quantitative prediction of peptide binding affinity by using hybrid fuzzy support vector regression," *Appl. Soft Comput.*, vol. 43, pp. 210–221, Jun. 2016.

[13] L. A. Zadeh, "The concept of a linguistic variable and its application to approximate reasoning—I," *Inf. Sci.*, vol. 8, no. 3, pp. 199–249, 1975.

[14] H. Hagras and C. Wagner, "Towards the wide spread use of type-2 fuzzy logic systems in real world applications," *IEEE Comput. Intell. Mag.*, vol. 7, no. 3, pp. 14–24, Aug. 2012.

[15] T. W. Chua and W. W. Tan, "Interval type-2 fuzzy system for ECG arrhythmic classification," in *Fuzzy Systems in Bioinformatics and Computational Biology*. Berlin, Germany: Springer, 2009, pp. 297–314.

[16] T. Nguyen, A. Khosravi, D. Creighton, and S. Nahavandi, "Medical data classification using interval type-2 fuzzy logic system and wavelets," *Appl. Soft Comput.*, vol. 30, pp. 812–822, May 2015.

[17] M. Zarinbal, M. H. F. Zarandi, I. B. Turksen, and M. Izadi, "A type-2 fuzzy image processing expert system for diagnosing brain tumors," *J. Med. Syst.*, vol. 39, no. 10, pp. 1–20, 2015.

[18] V. N. Vapnik, "An overview of statistical learning theory," *IEEE Trans. Neural Netw.*, vol. 10, no. 5, pp. 988–999, Sep. 1999.

[19] J. M. Mendel, R. I. John, and F. Liu, "Interval type-2 fuzzy logic systems made simple," *IEEE Trans. Fuzzy Syst.*, vol. 14, no. 6, pp. 808–821, Dec. 2006.

[20] T. Takagi and M. Sugeno, "Fuzzy identification of systems and its applications to modeling and control," *IEEE Trans. Syst., Man, Cybern.*, vol. SMC-15, no. 1, pp. 116–132, Jan./Feb. 1985.

[21] J. M. Mendel, *Uncertain Rule-Based Fuzzy Logic Systems: Introduction and New Directions*. Upper Saddle River, NJ, USA: Prentice-Hall, 2001.

[22] H. Wu and J. M. Mendel, "Introduction to uncertainty bounds and their use in the design of interval type-2 fuzzy logic systems," in *Proc. 10th IEEE Int. Conf. Fuzzy Syst.*, vol. 2, Dec. 2001, pp. 662–665.

[23] S. Greenfield, F. Chiclana, S. Coupland, and R. John, "The collapsing method of defuzzification for discretised interval type-2 fuzzy sets," *Inf. Sci.*, vol. 179, no. 13, pp. 2055–2069, 2009.

[24] M. Nie and W. W. Tan, "Towards an efficient type-reduction method for interval type-2 fuzzy logic systems," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jun. 2008, pp. 1425–1432.

[25] M. Biglarbegian, W. W. Melek, and J. M. Mendel, "On the stability of interval type-2 TSK fuzzy logic control systems," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 40, no. 3, pp. 798–818, Jun. 2010.

[26] C.-F. Juang, R.-B. Huang, and W.-Y. Cheng, "An interval type-2 fuzzy-neural network with support-vector regression for noisy regression problems," *IEEE Trans. Fuzzy Syst.*, vol. 18, no. 4, pp. 686–699, Aug. 2010.

[27] V. Uslan, H. Seker, and R. John, "A support vector-based interval type-2 fuzzy system," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2014, pp. 2396–2401.

[28] C. C. Chang and C. J. Lin, "LIBSVM: A library for support vector machines," *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, pp. 1–27, 2011.

[29] R. Nikhil, K. Pal, J. C. Bezdek, and T. A. Runkler, "Some issues in system identification using clustering," in *Proc. Int. Conf. Neural Netw.*, vol. 4, Jun. 1997, pp. 2524–2529.

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

[31] J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A k-means clustering algorithm," *Appl. Statist.*, vol. 28, no. 1, pp. 100–108, 1979.

[32] J. H. Ward, Jr., "Hierarchical grouping to optimize an objective function," *J. Amer. Statist. Assoc.*, vol. 58, no. 301, pp. 236–244, 1963.

[33] W. W. Tan, C. L. Foo, and T. W. Chua, "Type-2 fuzzy system for ECG arrhythmic classification," in *Proc. IEEE Int. Fuzzy Syst. Conf. (FUZZ-IEEE)*, Jul. 2007, pp. 1–6.

[34] B.-I. Choi and F. C.-H. Rhee, "Interval type-2 fuzzy membership function generation methods for pattern recognition," *Inf. Sci.*, vol. 179, no. 13, pp. 2102–2122, 2009.

[35] A. Stryhn, L. ø. Pedersen, A. Holm, and S. Buus, "Longer peptide can be accommodated in the MHC class I binding site by a protrusion mechanism," *Eur. J. Immunol.*, vol. 30, no. 11, pp. 3089–3099, 2000.

[36] W. W. P. Liao and J. W. Arthur, "Predicting peptide binding to major histocompatibility complex molecules," *Autoimmunity Rev.*, vol. 10, no. 8, pp. 469–473, Jun. 2011.

[37] I. A. Doytchinova, V. Walshe, P. Borrow, and D. R. Flower, "Towards the chemometric dissection of peptide–HLA-A 0201 binding affinity: Comparison of local and global QSAR models," *J. Comput.-Aided Mol. Des.*, vol. 19, no. 3, pp. 203–212, 2005.

[38] C. K. Hattotuwagama, P. Guan, I. A. Doytchinova, and D. R. Flower, "New horizons in mouse immunoinformatics: Reliable in silico prediction of mouse class I histocompatibility major complex peptide binding affinity," *Organic Biomol. Chem.*, vol. 2, no. 22, pp. 3274–3283, 2004.

[39] I. A. Doytchinova and D. R. Flower, "Physicochemical explanation of peptide binding to HLA-A 0201 major histocompatibility complex: A three-dimensional quantitative structure-activity relationship study," *Proteins, Struct., Function, Genet.*, vol. 48, no. 3, pp. 505–518, 2002.

[40] M. Gomez-Nunez *et al.*, "Peptide binding motif predictive algorithms correspond with experimental binding of leukemia vaccine candidate peptides to HLA-A 0201 molecules," *Leukemia Res.*, vol. 30, no. 10, pp. 1293–1298, 2006.

[41] S. Kawashima, H. Ogata, and M. Kanehisa, "AAindex: Amino acid index database," *Nucleic Acids Res.*, vol. 27, no. 1, pp. 368–369, 1999.

[42] C. Chrysostomou, H. Seker, and N. Aydin, "CISAPS: Complex informational spectrum for the analysis of protein sequences," *Adv. Bioinformatics*, vol. 2015, Dec. 2014, Art. no. 909765.

[43] Y. Saeys, I. Inza, and P. Larrañaga, "A review of feature selection techniques in bioinformatics," *Bioinformatics*, vol. 23, no. 19, pp. 2507–2517, 2007.

[44] D. Cai, C. Zhang, and X. He, "Unsupervised feature selection for multi-cluster data," in *Proc. 16th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, Jul. 2010, pp. 333–342.

[45] V. Uslan and H. Seker, "The quantitative prediction of HLA-B 2705 peptide binding affinities using support vector regression to gain insights into its role for the spondyloarthropathies," in *Proc. 37th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, Aug. 2015, pp. 7651–7654.

[46] F. Sarac, V. Uslan, H. Seker, and A. Bouridane, "Unsupervised selection of RV144 HIV vaccine-induced antibody features correlated to natural killer cell-mediated cytotoxic reactions," in *Proc. 38th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 3072–3075.

[47] V. Uslan and H. Seker, "Binding affinity prediction of S. Cerevisiae 14-3-3 and GYF peptide-recognition domains using support vector regression," in *Proc. 38th Annu. Int. Conf. Eng. Med. Biol. Soc. (EMBC)*, Aug. 2016, pp. 3445–3448.

[48] R. G. D. Steel, J. H. Torrie, and D. A. Dickey, *Principles and Procedures of Statistics*. New York, NY, USA: McGraw-Hill, 1997.

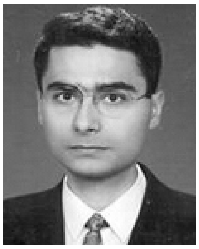
[49] J. L. Myers, A. D. Well, and R. F. Lorch, Jr., *Research Design and Statistical Analysis*, 2rd ed. Hillsdale, NJ, USA: Lawrence Erlbaum Associates, 2003.

[50] D. Wu and J. M. Mendel, "Designing practical interval type-2 fuzzy logic systems made simple," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2014, pp. 800–807.

[51] G. A. Miller, "The magical number seven, plus or minus two: Some limits on our capacity for processing information," *Psychol. Rev.*, vol. 63, no. 2, pp. 81–97, 1956.

[52] M. Almaraashi, R. John, A. Hopgood, and S. Ahmadi, "Learning of interval and general type-2 fuzzy logic systems using simulated annealing: Theory and practice," *Inf. Sci.*, vol. 360, pp. 21–42, Sep. 2016.

VOLKAN USLAN received the B.Sc. degree in computer engineering from Marmara University, istanbul, Turkey, the M.Sc. degree in computer engineering, and the Ph.D. degree in bioinformatics from De Montfort University, Leicester, U.K, where he was a part-time Lecturer and a Bioinformatics Researcher with the School of Computer Science and Informatics. His research interests include bioinformatics, computational genomics, data science, and machine learning.



HUSEYIN SEKER is currently a Multi-Disciplinary Data Scientist with particular interest in artificial intelligence, machine learning, extremely high-dimensional data in regression domain, and applications on both academia and industry. He is currently an Associate Professor/Reader with the Department of Computer and Information Sciences, Northumbria University, Newcastle upon Tyne, U.K. He is also the Director of the Institute of Coding at Northumbria University. He leads Smart Data Analytics Laboratory and co-lead the Information Management and Data Analytics Research Group. In addition to his academic duties, he is an Advisory Board Member of the North East Satellite Applications Centre of Excellence and Steering Group Member of Digital Catapult North East and Tees Valley. Further information can be found at http://computing.unn.ac.uk/staff/yqqd6/index.html.

ROBERT JOHN received the Degree (Hons.) in mathematics, the M.Sc. degree in statistics, and the Ph.D. degree in fuzzy logic. He worked in industry on achieving his M.Sc. in 1981 developing AI-based systems for industry. In 1989, he joined De Montfort University as a Lecturer in mathematics. He spent 24 years at De Montfort in various roles including heading a research group, the Head of Department, the Deputy Dean and the Head of research with the Faculty of Technology. He joined the University of Nottingham, in 2013 on the LANCS initiative and as the Head of the ASAP Research Group and is a member of LUCID. He was the Director of research with the School for 2 years, until 2018. He is currently the Associate Head of the School of Computer Science, University of Nottingham, Ningbo campus. He has authored more than 200 publications on fuzzy logic with more than 8500 citations.