Applied Soft Computing 116 (2022) 108280

Contents lists available at ScienceDirect

Applied Soft Computing

journal homepage: www.elsevier.com/locate/asoc

A novel ensemble fuzzy classification model in SARS-CoV-2 B-cell epitope identification for development of protein-based vaccine

Zeynep Banu Ozger [a,*], Pınar Cihan [b]

[a] Department of Computer Engineering, Sutcu Imam University, 46040, Kahramanmaras, Turkey
[b] Department of Computer Engineering, Tekirdag Namik Kemal University, 59860, Corlu, Tekirdag, Turkey

ARTICLE INFO

Article history:
Received 25 June 2021
Received in revised form 25 November 2021
Accepted 3 December 2021
Available online 15 December 2021

Keywords:
SARS-CoV-2
SARS-CoV
B-cell
Epitope
Spike protein
Fuzzy learning

ABSTRACT

B-cell epitope prediction research has received growing interest since the development of the first method. B-cell epitope identification with the aid of an accurate prediction method is one of the most important steps in epitope-based vaccine development, immunodiagnostic testing, antibody production, disease diagnosis, and treatment. Nevertheless, using experimental methods in epitope mapping is very time-consuming, costly, and labor-intensive. Therefore, although successful predictions with in silico methods are very important in epitope prediction, there are limited studies in this area. The aim of this study is to propose a new approach for successfully predicting B-cell epitopes for severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). In this study, the SARS-CoV B-cell epitope prediction performances of different fuzzy learning classification models genetic cooperative competitive learning (GCCL), fuzzy genetics-based machine learning (GBML), Chi's method (CHI), Ishibuchi's method with weight factor (W), structural learning algorithm on vague environment (SLAVE) and the state-of-the-art ensemble fuzzy classification model were compared. The obtained results showed that the proposed ensemble approach has the lowest error in SARS-CoV B-cell epitope estimation compared to the base fuzzy learners (average error rates; ensemble fuzzy=8.33, GCCL=30.42, GBML=23.82, CHI=29.17, W=46.25, and SLAVE=20.42). SARS-CoV and SARS-CoV-2 have high genome similarities. Therefore, the most successful method determined for SARS-CoV B-cell epitope prediction was used in SARS-CoV-2 cell epitope prediction. Finally, the eventual B-cell epitope prediction results obtained for SARS-CoV-2 with the ensemble fuzzy classification model were compared with the epitope sequences predicted by the BepiPred server and immunoinformatics studies in the literature for the same protein sequences according to VaxiJen 2.0 scores. We hope that the developed epitope prediction method will help design effective vaccines and drugs against future outbreaks of the coronavirus family, especially SARS-CoV-2 and its possible mutations.

1. Introduction

The immune system is a network of biological processes that protect an organism from various diseases. In organisms, it is responsible for preventing infections and eliminating established infections. There are 2 types of immune systems: innate and adaptive. Innate immunity is activated when an organism such as a bacteria or a virus enters the body, and since it has no immunological memory, it cannot recognize the same pathogen when it is encountered again. Adaptive immunity comes into play in situations where innate immunity is insufficient, such as viral infection, and since it contains immunological memory, it can recognize pathogens previously encountered, so it creates an immune response more quickly [1,2].

Antibodies in the blood and B/T white blood cells generate adaptive immune responses [3]. Because B/T lymphocyte cells contain memory, they are considered an important component of the adaptive immune system. These lymphocytes provide a protective function by producing antibodies. Antibodies are proteins that can recognize and bind biological substances called antigens. Each antibody has a compatible antigen and can only recognize it. In other words, individual antibodies are produced against each antigen [4]. B/T cells have specific receptors on their surface, and it is these receptors that enable them to recognize the antigen [5]. The part of antigens that binds to B/T cells or antibodies is called an epitope [4].

Epitopes are parts of the antigen that interact with B-cell receptors (BCRs). As seen in Fig. 1, B-cells are antibodies fighting against bacteria and viruses by creating a protein that is structured in a Y shape. B-cells recognize antigens using membrane-bound immunoglobulins (Igs). The antigen part of the B-cell that

[*] Corresponding author.
E-mail addresses: zeynepozger@ksu.edu.tr (Z.B. Ozger), pkaya@nku.edu.tr (P. Cihan).

https://doi.org/10.1016/j.asoc.2021.108280
1568-4946/© 2021 Elsevier B.V. All rights reserved.

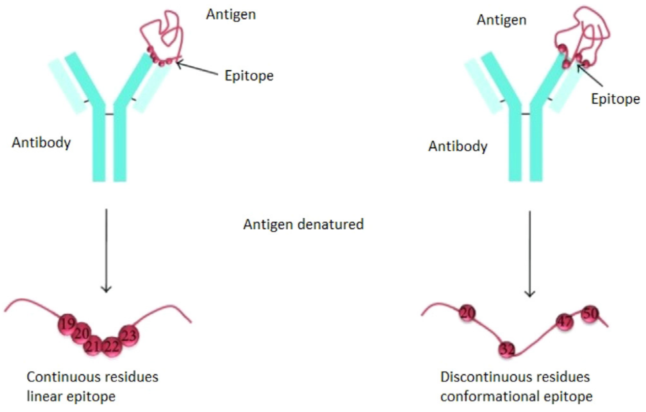

Fig. 1. Continues and discontinues B-Cell [7].

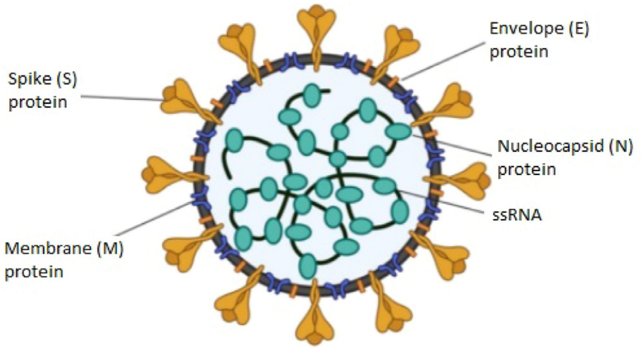

Fig. 2. Structure of SARS-CoV-2 [12].

binds to the immunoglobulin or antibody is called the B-cell epitope. When antigens bind to the antibody, the B-cell is activated and proliferates. Some of the proliferating cells form plasma cells, and some form memory cells. This immunological memory provides a fast and effective response to pathogens previously encountered [6].

B-cell epitopes are of 2 types, linear (continuous) and conformational (discontinuous) (Fig. 1). Although linear epitopes are relatively few, it is important to identify these epitopes, as they consist of peptides that can be easily used in antibody production. Continuous epitope prediction is both more convenient and easier to perform for antibody production [8].

Coronaviruses are single-stranded RNA viruses that are very common in animals. It contains 4 basic structural proteins. Of these, nucleocapsid (N), membrane (M), and envelope (E) proteins are responsible for the assembly of the virion. The spike (S) protein on the surface allows the virus to attach to the target cell [9]. Coronaviruses are divided into 4 main types: alpha, beta, gamma, and delta. Some alpha and beta species can cause respiratory tract infections in humans [10]. Severe acute respiratory syndrome coronavirus (SARS-CoV) and Middle East respiratory syndrome coronavirus (MERS-CoV) belong to the beta coronavirus family and have caused serious epidemic problems in the recent past. Due to its high similarity to SARS-CoV, this virus has been named SARS-CoV-2 by the Coronaviridae working group (CSG) of the International Committee on Virus Taxonomy [11]. The structure of SARS-CoV-2 is depicted in Fig. 2. The spike protein is the critical determinant of the SARS-CoV-2 genome, as it interacts with the host cell receptor.

Epitope data for SARS-CoV-2 are still limited, but since the gene and protein sequences of the virus are known, epitope information can be estimated by in silico analysis, taking into account past beta coronaviruses [13]. For this purpose, the sequence similarity of SARS-CoV-2 to other coronavirus species was examined. Considering the phylogenetic tree in Fig. 3, SARS-CoV-2 is more similar to the SARS-CoV virus than MERS-CoV [14].

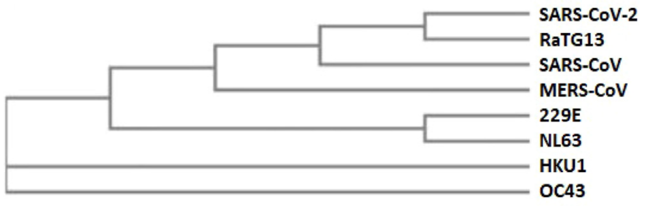

Fig. 3. Phylogenetic tree of beta coronaviruses [14].

The treatment of epidemics can be provided by community immunity or vaccination. Gaining community immunity requires a long process [15]. Since the epidemic causes loss of life and economic problems, it is important to develop an effective vaccine quickly. To develop a vaccine, it is necessary to identify protective immunogens against pathogens [16]. There is limited information as to which parts of the SARS-CoV-2 sequence are recognized by human immunity. This information is important for both vaccines and monitoring of mutations [17]. It is known that early immune responses against SARS-CoV-2 are mediated by IgM and IgA, while IgG responses are carried out 7–10 days after infection. It has been reported that developing neutralizing antibodies against the spike protein is important for an effective vaccine [18]. It has been found that directing antibodies to the receptor-binding domain and binding to spike trimers are important for long-term protective immunity against COVID-19 [19]. B-cell memory can reactivate antigen-specific responses upon re-exposure to infection [20]. Identifying conserved epitope regions would be useful in generating resistant immunity not only for the SARS-CoV-2 outbreak but also for ongoing virus evolution. Due to their strong immune responses, in vaccine studies developed for SARS-CoV and MERS, S, N, and M proteins were preferred as antigens [21–23].

Epitope identification with traditional methods is performed by experimental techniques, but this is a costly and time-consuming process [17]. For epitope identification by traditional methods, it is necessary to experimentally confirm by in vitro methods whether all possible subsequences in the protein sequence are antigens. With in silico approaches, subsequences that are less likely to be epitopes are eliminated using bioinformatics tools and historical data. Thus, the cost is reduced by reducing the number of epitope candidates that need to be examined by in vitro methods. [24]. Hereby, the determination of protein regions that are likely to be epitopes contributes to candidate vaccine and drug studies by narrowing the search space for epitopes. The aims of this study are to compare the prediction performance of fuzzy learning models for the prediction of epitope regions and to propose a new ensemble method for determining the epitope region by in silico analysis.

This study presents the following contributions and novelty:

- To determine SARS-CoV and SARS-CoV-2 B-cell epitope regions by in silico analyzes;
- To compare the predictive success of fuzzy learning models in identifying SARS-CoV B-cell epitope regions;
- To propose a novel ensemble approach to successfully predict SARS-CoV-2 B-cell epitope regions;
- To contribute to the development of new protein-based vaccines against SARS-CoV-2 and future epidemics by identifying epitope regions;

- To propose epitope candidates to assist biologists in developing a rapid and successful vaccine;
- It has been shown that more successful results are obtained with the proposed ensemble method compared to other studies in the literature;
- A statistically significant and robust ensemble approach has been proposed to identify SARS-CoV and SARS-CoV-2 B-cell epitope regions.

2. Related works

From the beginning of the SARS-CoV-2 epidemic, a lot of work has been done in this area and continues to be done. Most of the studies in this field are related to case/death estimation [25–28], detection of COVID-19 from medical images [29–31] or estimation of the number of vaccinated people [15]. There is a gap in the literature on the determination of epitope regions by in silico methods, which will be very beneficial to science and health services in both vaccine and drug development against the SARS-CoV-2 epidemics and future epidemics.

Authors in [32], in their study on epitope prediction, stated that 2 different ways were followed for epitope prediction by in silico analysis. The first of these is prediction methods based on SARS-CoV immunological data due to its genetic similarity, and the other is peptide binding prediction methods. The authors reviewed studies with both methods and predicted epitopes. Authors compared the epitopes obtained by in vitro methods with the epitopes estimated by in silico analysis, and they found that the methods using SARS-CoV immunological data, in general, coincided with the experimental results.

Within the scope of the study, B-cell epitope prediction was made for SARS-CoV-2 from SARS-CoV immunological data due to genetic similarity. For this reason, the prediction studies based on SARS-CoV immunological data have been examined within the scope of the literature. In this context, there are a limited number of studies in the literature. In some of the SARS-CoV-based studies, the sequence alignment results of SARS-CoV and SARS-CoV-2 were evaluated with bioinformatics tools to identify candidate epitopes. Some researchers have made predictions using the immunological data of SARS-CoV epitopes.

Nucleocapsid and spike proteins are the dominant structural proteins in the SARS-CoV-2 genome, as in other beta coronaviruses. In studies on the vaccine, it has been shown that spike protein is effective for developing a peptide vaccine and is a good candidate for generating a B-cell-dependent immune response [33]. In studies in which the nucleocapsid protein was experimentally tested for SARS-CoV, it was observed that it was the dominant protein expressed in the virion in the early stage of infection [34]. Studies have shown that the nucleocapsid protein is a strong T-cell-dependent immunogen [35]. Therefore, peptide-based studies on the immune response have focused on these 2 proteins. Within the scope of the study in [33], 34 linear B-cell epitopes, 29 MHC I, and 8 MHC II T-cell epitopes were shown as candidates for the vaccine.

Authors in [17] utilized the bioinformatics tools provided in the Immune Epitope Database (IEDB) and Virus Pathogen Resource (ViPR) to identify regions corresponding to SARS-CoV-2 sequences and to predict possible epitopes. IEDB is a database containing epitope information compiled from scientific literature for infectious disease, allergy, and autoimmunity. It also includes online bioinformatics tools to analyze epitope data and predict potential epitopes [36]. ViPR, on the other hand, is a database containing genome, gene, and protein sequence information about human pathogenic viruses [37]. In the related study, considering the conserved regions of SARS-CoV-2, B and T-cell epitope estimation for SARS-CoV-2 was realized based on sequence features. BepiPred 2.0 tool [4] was used for linear B-cell epitope prediction and Discotope 2.0 tool [38] for conformational B-cell epitope prediction. 29 epitopes for the spike protein, 4 for the nucleocapsid protein, and 3 for the membrane protein were identified as candidates.

In another study based on sequence features, Chen et al. [39] aimed to predict linear and conformational B and T-cell epitopes in the spike and nucleocapsid proteins of SARS-CoV-2. They identified the conserved regions of the virus genome by aligning the SARS-CoV-2 protein sequences obtained from the NCBI database with the Clustal Omega bioinformatics tool. Linear B-cell epitope prediction was performed with the BepiPred and ABCPred [40] tools and the epitope sequences with the highest antigenicity found were listed. Authors measured antigenicity values with the Vaxijen 2.0 [41] server. Conformational epitope prediction was performed with Discotope 2.0. Estimation of T-cell epitopes within the nucleocapsid protein that binds to the HLA-1 or HLA-2 molecule was made with the free online tool provided by IEDB. 63 B-cell epitopes have been proposed for vaccine studies.

At [42], authors performed B and T-cell epitope identification on spike protein. They used NetCTL 1.2 for T-cell epitopes, ElliPro and RaptorX for conformational B-cell epitopes, and BepiPred and ABCPred servers for linear B-cell epitopes. As a result of their analysis, they found 5 T-cell epitopes, 4 linear B-cell epitopes, and 5 conformational B-cell epitopes.

In the study [43], 115 T-cell epitopes and 298 B-cell epitopes were obtained from the NIAID and VIPR [37] databases, which were experimentally validated for SARS-CoV. These epitopes were aligned with the SARS-CoV-2 protein sequence and conserved and unmutated regions were identified. Accordingly, 27 T-cell epitopes and 42 linear B-cell epitopes for nucleocapsid and spike proteins, have been shown as candidates for SARS-CoV-2.

Sarkar et al. [44] identified possible B and T-cell epitopes using IEDB for spike, nucleocapsid, ORF3a, and membrane proteins. Among these epitopes, those with high antigenicity, non-allergenicity, and non-toxicity were identified as candidate epitopes for SARS-CoV-2. The authors suggested 5 epitopes for spike protein and 6 epitopes for nucleocapsid protein for vaccine studies.

The authors [45] predicted B and T-cell epitopes in spike, nucleocapsid, and membrane proteins of SARS-CoV-2 by an immunoinformatic method. Since B-cell epitopes can bind to antigen receptors on the B-cell surface, they eliminated intracellular epitopes from the epitopes they found with the BepiPred and BcePred servers. By measuring the antigenicity, allergenicity, and toxicity values for the remaining epitopes, they identified 10 B-cell linear epitopes with antigenicity greater than 0.9 as candidates.

In another immunoinformatics study [46], B-cell epitopes for spike protein were predicted with BepiPred 2.0. Those with a threshold value higher than 0.5 were also used for T-cell epitope prediction. Among the peptides found, the allergic and toxic ones were eliminated and 17 B-cell epitopes were presented as candidates.

Rehman et al. [47] focused on predicting immune response inducing epitopes in B and T-cells for multi-epitope vaccine design. Epitope prediction was performed with spike, Mpro, Nsp 12, and Nsp 13 proteins of SARS-CoV-2. As a result of the study, 46 antigenic B-cell peptides were predicted for the spike protein.

In [48], authors proposed a method to classify T-cell responses by analyzing TCR beta information from subjects infected and uninfected with SARS-CoV-2. The proposed method aimed to detect protective immunity acquired through natural infection or vaccine-induced immunity. Principal Component Analysis (PCA) and Hierarchical Clustering methods were applied to the sequence data separated into k-mers. Since the number of samples in the used dataset is small, the dataset is divided with hold-one-out. Accordingly, an accuracy value of 96% was obtained in the training data and 92.9% in the test data. The procedures were repeated for k-mers with a length of 3–9 amino acids, and the k-mers length with the highest success was determined as 5. The fact that the number of samples in the training dataset is too small has caused a situation that is overfitting to the training data. This situation reduces the generalizability of the proposed method.

Lee and Koohy [24] extracted T-cell peptides identified for SARS-CoV and peptides with high immunogenicity from IEDB. By aligning these peptides with those of SARS-CoV-2, they identified peptides with high sequence similarity as candidate peptides. MHC peptide connectivity of candidate peptides was measured with netMHCPan [49] and immunogenicity with iPred [50], high-value peptides are listed for vaccine studies.

Authors in [51] performed B-cell linear epitope prediction for SARS-CoV using an immunological epitope dataset [52] created with IEDB and UniProt. The authors made classification with Bayesian Neural Network, which is also used for uncertainty modeling in deep learning, with the thought that measuring uncertainty will also provide a measure for the reliability of the model. They achieved 85% accuracy in SARS-CoV data. Aleatoric and epistemic uncertainty methods were used to measure the uncertainty in epitope estimation. The related study was applied only for SARS-CoV epitope prediction, no prediction was made for SARS-CoV-2.

Noumi et al. [53] applied the Long Short Term Memory (LSTM) network with attention mechanism for epitope prediction in the IEDB dataset [52]. The results found were compared with the epitope sequences predicted by BepiPred 2.0 for the same protein sequences. The epitope peptide length is limited to 8–14 amino acids. The highest accuracy value was obtained as 0.79 for the case where the peptide length is 12.

In another study [54] on the IEDB epitope dataset [52], authors made epitope prediction for SARS-CoV by using immunological data with various machine learning methods. The authors used the dataset containing B-cell epitopes to develop the model and tested it with the SARS-CoV dataset. The most successful result was obtained with an accuracy of 87% with the ensemble learning model.

The coronavirus pandemic has proven that the World is not prepared for deadly viral outbreaks. Traditionally, it takes 15 or more years to develop a vaccine [55]. Thanks to in silico and computational methods, vaccine candidate epitopes can be successfully reduced, accelerating biologists in emergencies and epidemics. However, there is a gap in the literature on successful SARS-CoV-2 epitope prediction. The motivation of this study is to contribute to biologists in vaccine development by rapidly identifying a small number of vaccine candidate epitopes using in silico and bioinformatics tools.

3. Material and methods

In this study, we used the publicly available Kaggle dataset of SARS-CoV epitopes and SARS-CoV-2 peptides for the prediction of epitope regions. A novel ensemble fuzzy classification model was proposed for the successful prediction of epitope regions. R programming language was used for the development of fuzzy learning models and statistical analysis. Fuzzy rule-based classification systems (FRBCSs) belong to the soft computation family and are considered an effective approach to model complex problems. FRBCSs are specialized fuzzy rule-based systems and are used for handling classification problems. FRBCSs provide an interpretable model through the use of linguistic tags in their rules.

The general framework of the proposed model is formulated and given in Fig. 4. To train fuzzy methods, the labeled SARS-CoV dataset is used. To get statistical validity, the dataset was divided into train and test sets 6 times using random sampling with replacement. The fuzzy rule sets obtained during the training phase were applied to the test sets and their performances were measured. Five fuzzy methods (GBML, GCCL, CHI, SLAVE, W) were applied to all training sets separately and the final decisions for the relevant test set were made by the majority voting method in the individual decisions of these 5 methods. In this way, it is obtained one ensemble model for each train set. The prediction was made using all models with unlabeled SARS-CoV-2 data, and the class of each peptide was obtained by combining the decisions of ensemble methods.

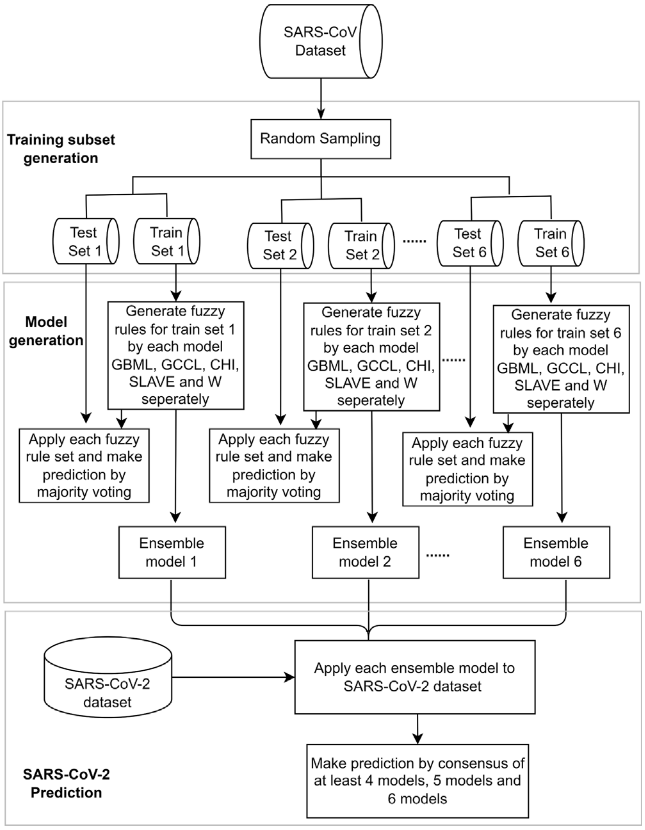
SARS-CoV Dataset
Training subset generation
Random Sampling
Test Set 1
Train Set 1
Test Set 2
Train Set 2
......
Test Set 6
Train Set 6
Model generation
Generate fuzzy rules for train set 1 by each model GBML, GCCL, CHI, SLAVE and W seperately
Generate fuzzy rules for train set 2 by each model GBML, GCCL, CHI, SLAVE and W seperately
......
Generate fuzzy rules for train set 6 by each model GBML, GCCL, CHI, SLAVE and W seperately
Apply each fuzzy rule set and make prediction by majority voting
Apply each fuzzy rule set and make prediction by majority voting
Apply each fuzzy rule set and make prediction by majority voting
Ensemble model 1
Ensemble model 2
......
Ensemble model 6
SARS-CoV-2 dataset
Apply each ensemble model to SARS-CoV-2 dataset
Make prediction by consensus of at least 4 models, 5 models and 6 models
SARS-CoV-2 Prediction

Fig. 4. General framework of the ensemble fuzzy classification model.

The datasets used in this study are described in Section 3.1, the FRBCS methods are briefly explained in Section 3.2, the proposed model is given in Section 3.3, and the evaluation metrics are described in Section 3.4.

3.1. Dataset description

In this study, a dataset [52] containing B-cell epitopes obtained from IEDB and UniProt was used to predict SARS-CoV-2 epitopes. There are 3 datasets here: B-cell epitopes, SARS-CoV epitopes and SARS-CoV-2 peptides. Of these, B-cell epitopes and SARS-CoV epitopes are labeled data, and the SARS-CoV-2 dataset contains peptides of various lengths that are identified from a protein sequence and have no label information. In this study, due to the high genome similarity, the SARS-CoV epitope dataset was used to develop a fuzzy model. The model with the high test set accuracy for SARS-CoV was also applied to the SARS-CoV-2 dataset, and epitope prediction was made.

The datasets contain 13 features. The SARS-CoV and B-cell datasets also have target values, indicating whether an amino

Table 1
Peptid and protein based features at datasets.

Feature	Description
Chou–Fasman	Peptide feat. Relative frequency analysis on the basis of amino acids for tertiary structural elements. Given here for B-Turn.
Emini	Peptide feat, relative surface accessibility, a measure of residue solvent exposure.
Kolaskar–Tongaonkar	Peptide feat. Antigenicity, antigenic propensity of residues.
Parker	Peptide feat. A measure of hydrophobicity of peptide.
Isoelectric-point	Protein feat. pH value of the amino acid in an electric field.
Aromaticity	Protein feat. A factor for protein fragment solubility.
Hydrophobicity	Protein feat. A measure of the degree of affinity between water and the side chain of an amino acid.
Stability	Protein feature.

acid peptide is capable of inducing antibodies. The proteins in the datasets are immunoglobulin antibody proteins, as they are the most common type of antibody found in the bloodstream. The dataset includes protein and peptide sequences, protein IDs, starting and ending positions of peptides in the protein sequence, and protein/peptide-based features. B-cell epitope prediction is based on the antigenicity, hydrophobicity, surface accessibility, beta turns, and flexibility properties of epitopes. The features in the dataset and their descriptions are given in Table 1.

There are a total of 520 samples in the SARS-CoV dataset. A total of 140 of them are in the positive class, and the remaining 380 samples are in the negative class. Positive class means that the corresponding peptide is the epitope. The longest peptide is 393 amino acids long, and the shortest peptide is 5 amino acids long. The SARS-CoV-2 dataset includes 20312 samples, and the peptides are 5–20 amino acids long.

3.2. Fuzzy learning classification models

The genetic cooperative competitive learning (GCCL) [56] algorithm uses genetic cooperative competitive learning to handle classification problems. In this technique, a chromosome defines each linguistic IF–THEN rule using integers as the representation of the previous part. The heuristic is applied to automatically produce the class in the consequence part of the fuzzy rules. Assessment is calculated separately for each rule. Thus, performance is not based on the whole rule set.

The fuzzy genetics-based machine learning (GBML) model [57] is based on a hybridization of Ishibuchi's genetic collaborative competitive learning (GCCL) and Pittsburgh approaches. Selection, crossover, and mutation operators of the genetic algorithm are applied according to the algorithm proposed by Pittsburgh. Here, each rule set is treated as an individual. Then, GCCL steps are applied to each of the created rule sets with a probability specified as a parameter in the algorithm. Good fuzzy rules are found efficiently with the GCCL approach.

Chi's (CHI) method [58] is proposed to overcome classification problems and is an extension of Wang and Mendel method. This algorithm is similar to the technique of Wang and Mendel's [59]. Chi's method generates fuzzy IF–THEN rules and then replaces them with class labels so that they are sequential parts. Regarding the calculation of the degrees of each rule, they are identified by the previous (antecedent) part of the rules. Redundant rules can be eliminated according to their degree. Thus, fuzzy IF–THEN rules based on the FRBs model are obtained. Calculation of the degree of each rule is determined by the antecedent part of the rules.

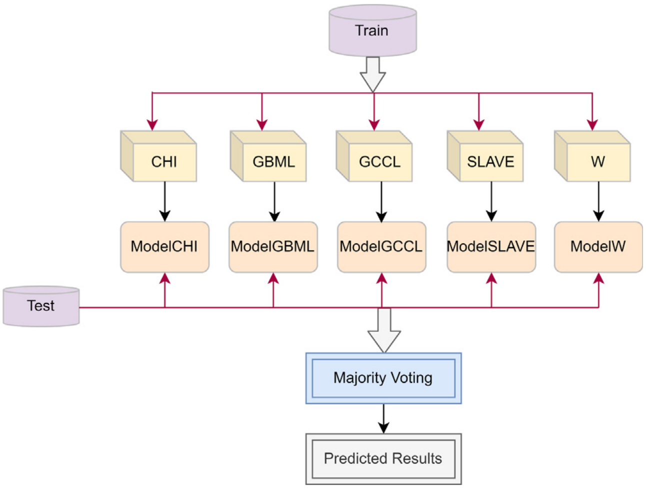

Fig. 5. The proposed ensemble fuzzy classification model.

Ishibuchi's method with a weight factor (W) applies the second type of FRBs, which has weights in consequent parts of the rules [60]. The antecedent fragments are then determined from the training data by a grid-type fuzzy partition. The resulting class is defined as the dominant class in the fuzzy subspace corresponding to the antecedent part of each fuzzy IF–THEN rule. The class of a new instance is determined by the rule's resulting class, which is the maximum product of its compatibility and precision. The degree of concordance is determined by summing the degrees of the membership functions of the previous sections, while the degree of precision is calculated from the ratio between the next class.

The structural learning algorithm in a vague environment (SLAVE) is based on an approach where only one fuzzy rule is obtained each time the genetic algorithm is run. To remove unrelated variables in a rule, SLAVE has a two-part structure: the first part demonstrates the relevance of the variables, and the second part describes the values of the parameters. This method applies binary codes as representative of the population and executes basic genetic operators, i.e., crossing, selection, and mutation on the population. Then, the best rule is identified as the rule with the highest degree of integrity and consistency [61].

3.3. Proposed ensemble fuzzy classification approach

An ensemble fuzzy classifier technique is proposed to develop models and make predictions on the SARS-CoV dataset. Five different fuzzy methods were applied separately to the training data created by random sample selection. The proposed ensemble model combines the decisions classifiers GCCL, GBML, CHI, W, and SLAVE by using a majority voting scheme. As shown in Fig. 5, with the model developed with each of them, predictions were made on the test dataset consisting of random samples. By combining the decisions of the models, the class of each sample in the test set was decided by the majority voting method.

By random sampling with replacement, the training and test set creation process was repeated 6 times. The performance of the developed system was measured by applying the proposed ensemble fuzzy model to each training-test dataset and taking the average.

Based on the high genome similarity of SARS-CoV and SARS-CoV-2, a fuzzy model was created with SARS-CoV data, and epitope prediction was made by giving unlabeled SARS-CoV-2 data

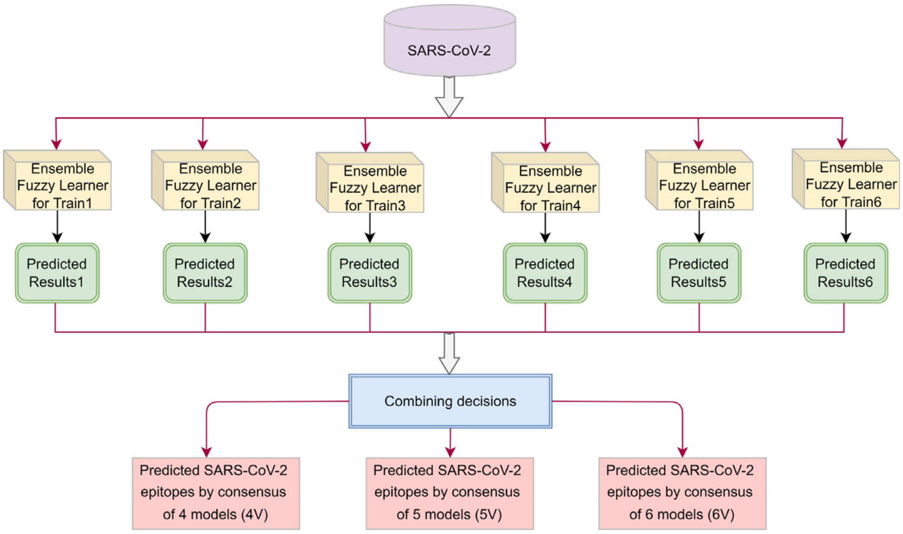
SARS-CoV-2
Ensemble Fuzzy Learner for Train1
Ensemble Fuzzy Learner for Train2
Ensemble Fuzzy Learner for Train3
Ensemble Fuzzy Learner for Train4
Ensemble Fuzzy Learner for Train5
Ensemble Fuzzy Learner for Train6
Predicted Results1
Predicted Results2
Predicted Results3
Predicted Results4
Predicted Results5
Predicted Results6
Combining decisions
Predicted SARS-CoV-2 epitopes by consensus of 4 models (4V)
Predicted SARS-CoV-2 epitopes by consensus of 5 models (5V)
Predicted SARS-CoV-2 epitopes by consensus of 6 models (6V)

Fig. 6. Prediction process for SARS-CoV-2.

Table 2
The parameters of fuzzy models.

Parameter	Description	Value
popu.size	Population size	GCCL:30, GBML:10
num.class	Number of classes	For all methods:2
num.labels	Number of linguistic terms	W:11, CHI:5, GCCL:9, GBML:7, SLAVE:7
persen_cross	Probability of crossover	GCCL:0.8, GBML:0.9, SLAVE:0.8
persen_mutant	Probability of mutation	GCCL:0.4, GBML:0.2, SLAVE:0.4
max.gen	Maximum number of generations	GCCL:150, GBML:10, SLAVE:40
type.mf	The type of the shape of the membership function	W:Gaussian, CHI:Triangle
type.tnorm	The type of the tnorm	W:min, CHI:min
type.snorm	The type of the snorm	W:sum, CHI:max
type. implication. func	Type of implication functions	W:Dienes Recher, CHI:Zadeh
max.num.rule	Maximum number of rules	GBML:10
p.dcare	A probability of “don’t care” attributes occurred	GBML:0.5
p.gccl	A probability of GCCL process occurred	GBML:0.4
max.iter	Maximum number of iterations	SLAVE:30
k.lower	A lower bound of the noise threshold	SLAVE:0
k.upper	A value between 0 and 1 representing the level of generalization	SLAVE:0.8

to these models as test data. Since the fuzzy methods used are heuristic, 6 training-test sets were created by random sampling from all SARS-CoV data, and an ensemble model was obtained by training each training set with five different fuzzy methods (modelChi, ModelGBML, modelGCCL, modelSlave, modelW in Fig. 5). SARS-CoV prediction successes were measured by majority voting for each training-test set pair. Since SARS-CoV data were divided into 6 training-test sets, epitope prediction was made by applying six models consisting of SARS-CoV-trained models of five fuzzy methods to SARS-CoV-2 data. Each yellow box in Fig. 6 contains the training and model building processes shown in Fig. 5. The decisions made by the models are combined with different degrees of precision. The final epitope decision-making strategies were named 4V (at least 4 votes), 5V (at least 5 votes), and 6V (at least 6 votes). The epitope prediction process for SARS-CoV-2 is shown in Fig. 6.

The details of the parameter settings for each model are given in Table 2. All specified parameters were determined experimentally. In the proposed method, the labeled SARS-CoV dataset was used to develop a model with fuzzy methods. Since the problem under consideration is a classification problem, the degrees of the rules are determined in all methods depending on how much they represent the data during training. That is, the degree of membership is directly proportional to the fact that the rule represents the training data. Membership functions are defined with the ‘type.mf’ parameter specified in Table 2. The main difference between the methods is the way in which the learning and fuzzy rules are created.

3.4. Evaluation metrics

The epitope prediction performances of the models on the SARS-CoV dataset were compared according to accuracy rate, error rate, sensitivity/recall rate (RR), specificity rate (SR), positive predictive value (PPV) and negative predictive value (NPV) criteria. These metrics are calculated from the confusion matrix. The confusion matrix or contingency table summarizes the performance of a classification model. The accuracy rate is the ratio of all correctly predicted epitopes to the total number of epitopes. The error rate is the ratio of all incorrectly predicted epitopes to the total number of epitopes. The sensitivity or recall rate (RR) metric is a measure of how well a test identifies true positives. RR is the ratio of the true epitopes over all the actual epitopes. The SR is the ratio of the true non-epitopes over all the actual non-epitopes. PPV is the ratio of all the true epitopes over all the predicted epitopes. NPV is the ratio of the true non-epitopes over all the predicted non-epitopes. These metrics are expressed mathematically as follows:

$$Accuracy\ rate = \frac{TP + TN}{TP + TN + FP + FN} * 100 \tag{1}$$

$$Error\ rate = \frac{FP + FN}{TP + TN + FP + FN} * 100 \tag{2}$$

$$Sensitivity\ rate = \frac{TP}{TP + FN} * 100 \tag{3}$$

$$Specificity\ rate = \frac{TN}{TP + FP} * 100 \tag{4}$$

$$PPV = \frac{TP}{TP + FP} * 100 \tag{5}$$

$$NPV = \frac{TN}{TN + FN} * 100 \tag{6}$$

In the equations, TP, TN, FP, and FN represent true positives, true negatives, false positives, and false negatives, respectively.

Furthermore, the eventual B-cell epitope prediction results obtained for SARS-CoV-2 were compared with the epitope sequences predicted by the BepiPred server for the same protein sequences. BepiPred is a web server that predicts B-cell epitopes from antigen sequences (http://www.cbs.dtu.dk/services/BepiPred/). BepiPred makes predictions with a model trained using Random Forest on a dataset of 649 antigen–antibody crystal structures.

The Vaxijen server was used to compare the SARS-CoV-2 epitopes predicted by the BepiPred server and presented in the literature with the epitopes found by the proposed method. (http://www.ddg-pharmfac.net/vaxijen/VaxiJen/VaxiJen.html). This antigenicity measurement tool is used to analyze B and T-cell epitopes by evaluating the physical and chemical properties of amino acids and their abundance in known B and T-cell epitopes. The higher the epitope antigenic score, the more likely it is to be used as an antigen, i.e., it has greater immunogenicity.

4. Experimental results

4.1. Dataset preprocessing

In the SARS-CoV and SARS CoV-2 datasets, different peptides were identified for a single IgG protein with lengths of 1255 and 1281 amino acids, respectively. Therefore, parent protein ID and protein-based features were the same for all data, so these features were excluded from both datasets. Additionally, the features that give the start and end positions of the peptide and peptide sequence features were also removed, and a new feature including the peptide length was added. As a result, the datasets were arranged to contain a total of 5 features. There is also a label feature for SARS-CoV. The correlation matrix of independent variables of the SARS-CoV dataset, density plots, and 2D density charts are given in Fig. 7.

In Fig. 7, the lower triangle shows the 2D density of the combination between the two variables. The Pearson correlation is given on the upper triangle, and the variable distributions are illustrated on the diagonal. Linear dependence between two variables was measured with Pearson correlation. In the upper triangle, both the correlation coefficient and the correlation significance level are given (***$P < 0.001$, **$P < 0.01$, *$P < 0.05$). When the scatter plots are examined, it is seen that the variables are normally distributed and the highest and most significant correlation is between Parker and Chou–Fasman features ($R = 0.67$, $P < 0.001$).

4.2. SARS-CoV prediction

SARS-CoV dataset is unbalanced in terms of label distribution. Of the 520 samples in the dataset, 140 are in the positive (epitope) class, while the remaining 380 are in the negative (non-epitope) class. The training, and test sets are divided according to the class information. Out of 140 samples in the positive class, 120 samples were randomly selected for training so that the model could learn the data and the remaining 20 samples were used for testing. Since there were few samples in the positive class, it was observed that the model could not learn the positive class when the number of samples included in the test set was increased. Therefore, 20 samples were randomly selected from the negative class so that there were equal numbers of samples from both classes in the test set.

Table 3
Classification errors for SARS-CoV.

Method	Test1	Test2	Test3	Test4	Test5	Test6	Avg.	Sig
CHI	17.5	37.5	25	27.5	22.5	20	29.17	+
GBML	17.5	15	32.5	15	20	22.5	23.82	+
GCCL	22.5	55	30	12.5	30	32.5	30.42	+
SLAVE	27.5	17.5	7.5	25	22.5	22.5	20.42	+
W	52.5	52.5	52.5	27.5	45	47.5	46.25	+
Ensemble fuzzy	7.5	12.5	7.5	10	7.5	5	8.33	

For the model to learn the classes correctly, it was decided experimentally how many samples from the negative class should be present in the training set. There were 120 samples from the positive class in the training set. If there are 360, 300, 240, 180, and 120 samples from the negative class, the estimation error according to the classes and the total estimation error of the proposed ensemble fuzzy model are given in Fig. 8.

As shown in Fig. 8, when the number of samples for the negative class was higher than that for the positive class, high prediction accuracy was obtained for the negative class, but the model could not recognize the positive class. In the training set, as the number of samples started to be equally distributed according to the classes, the model's ability to correctly predict the positive class increased, and the total error decreased. From this point of view, the training set was created to have 120 samples from both classes.

The training-test set creation process was repeated 6 times to include random samples. Accordingly, 20 randomly selected out of 140 samples in the positive class were allocated as testing, and the rest were allocated as training. For the negative class, 20 randomly selected out of 380 samples were added to the test set, and 120 randomly selected samples were added to the training set.

In Table 3, the individual decisions of fuzzy methods and the prediction errors obtained by the proposed method are given for the test sets. When the individual decisions of fuzzy methods are examined, they are insufficient on their own for defining membership functions that can model the whole data. It is clear that combining the decisions of fuzzy methods has resulted in a significant improvement in prediction performance. This is because each method learns different properties in the data. The proposed ensemble fuzzy model classifies SARS-CoV data with an average accuracy of 91.7%. The Wilcoxon rank-sum test was applied to SARS-CoV results to measure the statistical significance of the difference between the individual methods and the proposed method results. The test results are given in the last column of Table 3 for $\alpha = 0.05$. '+' indicates that the results of the proposed method are statistically better than those of the corresponding algorithm.

4.3. SARS-CoV-2 prediction

Considering the high genome similarity of SARS-CoV with SARS-CoV-2, epitope prediction was made for SARS-CoV-2 with

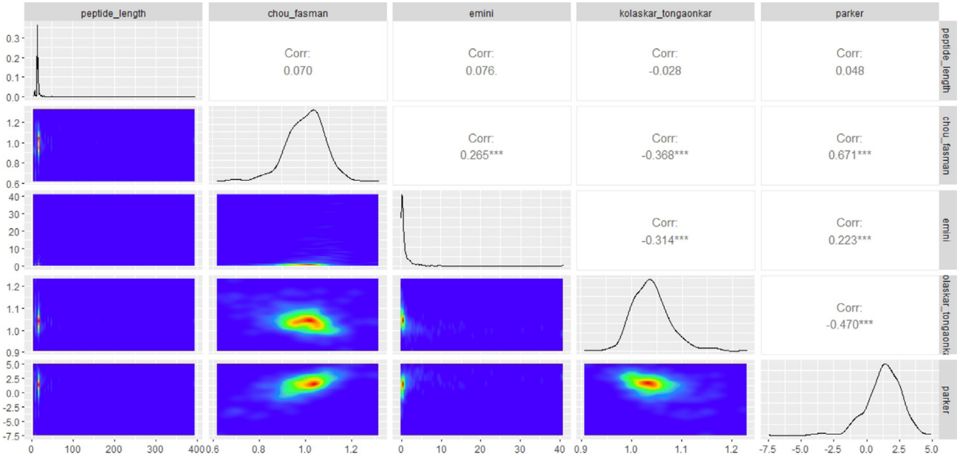

Fig. 7. Correlation, density and 2D density plot of independent variables.

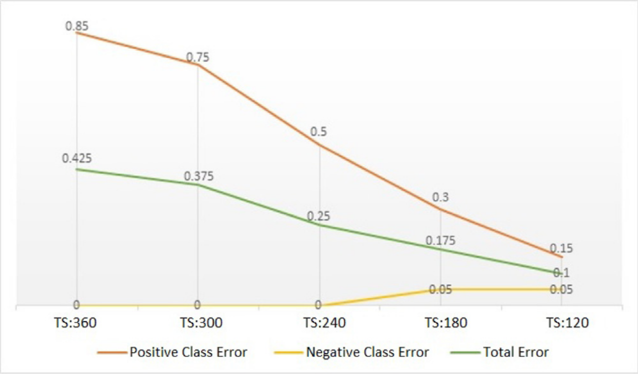

Fig. 8. Train size tuning for negative class samples.

fuzzy models trained for SARS-CoV, as shown in Fig. 6. While estimating with SARS-CoV data, the model was trained 6 times since there were 6 training-test sets created with randomly selected samples. Each of these models was also applied to the SARS-CoV-2 data to make predictions.

The ensemble fuzzy classification method makes predictions by combining the decisions of 5 fuzzy classifiers. This process was repeated 6 times to ensure statistical validity. Therefore, 6 models were formed. With each model, the prediction was made on all SARS-CoV-2 data. Unlike the method applied for SARS-CoV, the decision of the models is combined with different degrees of sensitivity; common decision of at least 4 models (4V), common decision of at least 5 models (5V) and common decision of all models (6V). Accordingly, for a peptide in the dataset to be labeled as an epitope by the 4V method, at least 4 out of 6 models must have made an "epitope" decision for that peptide. Table 4 gives the number of peptides labeled as epitopes for each method and their lengths. Additionally, the "dataset" column shows how many peptides the data include for each length.

The SARS-CoV-2 dataset contains all possible k-mers of the spike protein that are 5–20 amino acids long. The proposed ensemble fuzzy classification model labeled 5465 peptides with the 4V method, 3911 peptides with the 5V method, and 2004 peptides with the 6V method as the epitope. The predicted epitopes for all three methods are listed in Appendix Table A.1.

The algorithm was executed on an Intel(R) Core (TM) i7-6700 HQ CPU at 2.60 GHz, on a 64-bit architecture with 16 GB RAM, running Windows 10 and the R programming language using the frbs package. The execution time results are given in Table 5.

Table 4
Prediction results for SARS-CoV-2.

Epitope	Number of predicted epitopes			
Length	Dataset	4V	5V	6V
5	1277	776	642	325
6	1276	201	136	70
7	1275	229	157	67
8	1274	265	185	81
9	1273	287	194	88
10	1272	294	196	82
11	1271	313	206	106
12	1270	330	219	98
13	1269	321	226	123
14	1268	321	232	121
15	1267	329	221	121
16	1266	335	229	138
17	1265	345	244	145
18	1264	369	273	144
19	1263	369	269	139
20	1262	381	282	156
Total	20 312	5465	3911	2004

Table 5
CPU time for SARS-CoV-2 prediction.

TrainSet	GCCL (min)	W (s)	CHI (s)	GBML (min)	SLAVE (min)	Prediction (min)	Total (min)
Train1	1.73	0.07	0.03	4.25	4.57	9.55	20.1
Train2	1.78	0.06	0.03	4.33	4.61	9.08	19.8
Train3	1.44	0.06	0.04	4.31	4.34	8.92	19.33
Train4	1.80	0.06	0.03	4.28	4.54	9.07	19.69
Train5	1.86	0.06	0.03	4.35	4.46	9.01	19.68
Train6	1.84	0.06	0.03	4.27	4.49	9.07	19.67
CPU time of whole framework							118.27

As mentioned earlier, the SARS-CoV dataset was divided into 6 training sets by random sampling. The training model in each row represents execution time for related training set for all methods. A separate model was created for all fuzzy methods in each training set. All models obtained with a training set were estimated by giving SARS-CoV-2 data as a test set. For this reason, model creation and prediction times are given separately for each training set. The 'prediction' column is the time required to make predictions in the SARS-CoV-2 data for models trained with the relevant training set. The last column is the time required to train model with relevant training set and to make prediction. The time required to train all models and make predictions for all training sets is given in the last line. The final decision for the SARS-CoV-2 data was obtained by estimating all models from all training

Table 6
Comparison with BepiPred results.

BepiPred			4V			5V			6V		
Epitope	Len	Ant	Det	Len	Ant	Det	Len	Ant	Det	Len	Ant
VNLTTRTQLPPAYTNSFTR	19	**0.6285**	✓	19	**0.6285**	✓	19	**0.6285**	✗	–	–
ASTEKS	6	0.6206	✓	7	**0.8587**	✓	7	**0.8587**	✓	7	**0.8587**
PFLGVYYHKNNKSWMESE	18	**0.5664**	✓	18	**0.5664**	✓	18	**0.5664**	✓	18	**0.5664**
KHTPINLVRDLPQGFSA	17	**0.6207**	✓	19	0.5535	✗	–	–	✗	–	–
TPGDSSSGWTA	11	0.2473	✓	12	0.0746	✓	17	**0.4892**	✓	17	**0.4892**
IYQTSNFRVQP	11	**1.0147**	✓	12	0.9986	✓	12	0.9986	✓	16	0.8559
DEVRQIAPGQTGKIAD	16	1.0388	✓	16	1.0388	✓	16	1.0388	✓	19	**1.1515**
NNLDSKVGGNYN	12	**0.7538**	✓	15	0.7275	✓	15	0.7275	✓	15	0.7275
GFNCYFPLQSYGF	13	0.8519	✓	18	**0.8567**	✓	18	**0.8567**	✓	18	**0.8567**
SNKKFLPF	8	**1.3952**	✓	8	**1.3952**	✓	8	**1.3952**	✓	9	1.1432
NCTEV	5	NA	✓	5	NA	✓	5	NA	✗	–	–
HADQLTPT	8	0.4177	✓	8	0.4177	✓	8	0.4177	✓	16	**0.6093**
RVYSTGSNVFQ	11	−0.1000	✓	13	**0.3359**	✓	13	**0.3359**	✓	14	0.1826
AYTMSLGAENSVAYSNN	17	**0.5966**	✓	17	**0.5966**	✓	17	**0.5966**	✓	17	**0.5966**
KQIYKTPPIKDFGGF	15	−**0.3896**	✓	15	−**0.3896**	✓	15	−**0.3896**	✓	15	−**0.3896**
LPDPSKPSKR	10	**0.2641**	✓	10	**0.2641**	✓	10	**0.2641**	✓	10	**0.2641**
DPPEAEVQI	9	**0.5966**	✓	10	0.4955	✓	10	0.4955	✓	11	−0.0004
GQSKRVDFC	9	**1.7790**	✓	11	1.4088	✓	12	1.3607	✓	12	1.3607
FYEPQIITTD	10	0.4179	✓	10	0.4179	✓	16	**0.6504**	✓	19	0.2751
VNNTVYDPLQPELDSF	16	**0.2201**	✓	16	**0.2201**	✓	16	**0.2201**	✓	19	0.1493
LGKYEQYIKGSGR	13	**0.3101**	✓	13	**0.3101**	✓	13	**0.3101**	✓	13	**0.3101**
Average		0.5925			0.5900			0.6253			0.5553

sets. Since GCCL, GBML and SLAVE are genetic algorithm-based iterative methods, their run times are greater than those of CHI and W.

Since SARS-CoV-2 is unlabeled, the selected epitopes were compared with the epitopes that the BepiPred server found for the same protein. In addition, these results have been compared with epitopes found in studies with various bioinformatics tools or in vitro methods in the literature. The BepiPred server identified 44 peptides for the spike protein, 1–36 amino acids long. For the same protein sequence, in [33] 34, in [17] 29, in [39] 63, in [42] 4, in [43] 21, in [44] 5, in [46] 17, in [47] 46, and in [45] 10 linear B-cell epitopes were identified as vaccine candidates.

The peptides in the SARS-CoV-2 dataset are 5–20 amino acids long. Epitopes shorter than 5 amino acids or longer than 20 amino acids among the epitopes compared in BepiPred and the literature were not included in the comparison. In addition, some peptides identified in these studies were not included in the comparison because they were not included in the SARS-CoV-2 dataset used.

After elimination, comparisons were made for different sensitivities (4V, 5V, 6V) of the proposed method. Comparative results for BepiPred are given in Table 6, and comparative results for the literature are given in Table 7. The first column in the tables includes peptides BepiPred or found in studies in the literature. Other columns are the sequence lengths of those peptides (Len) and antigenicity scores (Ant) measured by Vaxigen 2.0. Comparison results, peptide lengths and antigenicity scores of the proposed method for different sensitivity levels are given in the next columns. The Detection column (Det) indicates whether a peptide is found by the proposed method. A peptide identified by related studies is also marked "✓" if it is a subsequence of a peptide found by the ensemble fuzzy method. Those with high antigenicity scores are written in bold. The mean antigenicity values of the peptides found by the methods are given in the "Average" line. If the antigenicity score of a peptide could not be measured with the Vaxigen tool, it is indicated as "NA".

As seen from Table 6, 21 of the sequences identified by BepiPred were found by the 4V method, 20 by the 5V method, and 18 by the 6V method. When peptides with different sequence lengths were compared according to their antigenicity scores, 6 peptides found by the ensemble fuzzy method had higher antigenicity scores, while BepiPred was more successful for 5 peptides. Looking at the average scores, the 5V method gives the best result.

In Table 7, epitopes in other studies in the literature are compared with the proposed method. Grifoni et al. [17] identified 19 peptides as candidates for the vaccine. Of these, 12 were also estimated by the 4V and 5V methods, and 10 by the 6V methods. Considering all 3 methods, the 8 predicted peptides appeared to be more antigenic. According to the mean antigenicity score, the average of all peptides found by the 6V method was the highest.

All 33 peptides identified by [39] were also predicted by the 4V method, and the 5V and 6V methods selected 30 and 22 of them, respectively. Of the peptides predicted by the proposed method, 13 have higher antigenicity scores. The peptides with the mean highest antigenicity scores are the peptides predicted by the 5V method.

In [44], the authors identified 4 peptides in their study, all of them which were selected by the ensemble fuzzy method, and the developed method determined more antigen epitopes for 2 of them.

Of the 16 epitopes identified in [43], 9, 6, and 5 epitopes were labeled by the 4V, 5V, and 6V methods, respectively. Except for 2, the developed method selected more antigenic epitopes. In terms of the mean antigenicity score, the 5V method was the most successful.

Eighteen B-cell vaccine candidates were identified by [33], of which 16 were nominated by the 4V and 5V methods, and 11 were nominated for the vaccine by the 6V method. Considering the antigenicity of peptides of different lengths, more antigen epitopes are estimated for the 4 vaccine candidates identified by the developed method.

Of the 42 B-cell epitopes identified by Rehman et al. [47], 34 were detected by the 4V method, 24 by the 5V method and 13 by the 6V method. The antigenicity scores of 15 of the epitopes found in different lengths by the ensemble fuzzy method were higher.

Lin et al. [45] identified 4 epitopes in their study, 3 of which were found by the fuzzy method at all sensitivity levels. However, the epitopes found by [45] seemed to have higher average antigenicity scores.

All 7 peptides presented in the study [46] were predicted by the 4V and 5V methods, but the 6V method detected 3 of

Table 7
Comparison results with literature.

Predicted Epitopes in [17]			4V			5V			6V		
Epitope	Len	Ant	Det	Len	Ant	Det	Len	Ant	Det	Len	Ant
FHAIHVSGTNG	11	**0.8882**	✓	18	0.6317	✓	18	0.6317	✓	18	0.6317
TLDSKTQSLLIVNNATNV	18	**0.7295**	✗	–	–	✗	–	–	✗	–	–
PGDSSSGWTAGA	12	0.0820	✓	17	0.2386	✓	17	0.2386	✓	18	**0.5144**
NENGTITDA	9	0.5257	✓	9	0.5257	✓	10	0.6020	✓	11	**0.7882**
IYQTSNFRV	9	0.3109	✓	11	0.2839	✓	11	0.2839	✓	16	**0.8559**
IAWNSNNLDSK	11	**1.2444**	✓	11	**1.2444**	✓	12	0.9178	✓	13	0.7773
STEIYQAGSTPCNGV	15	−0.0513	✓	16	**0.0539**	✓	18	−0.0751	✓	19	−0.0745
RVYSTGSNVFQTRA	14	0.3248	✓	14	0.3248	✓	14	0.3248	✓	18	**0.5620**
GAEHVNNSYE	10	**0.8739**	✓	10	**0.8739**	✓	10	**0.8739**	✓	10	**0.8739**
YICGDSTECSNLLLQ	15	−**0.0093**	✗	–	–	✗	–	–	✗	–	–
GSFCTQLNRALTG	13	**0.4763**	✗	–	–	✗	–	–	✗	–	–
AVEQDKNTQE	10	0.2792	✓	12	**0.5008**	✓	12	**0.5008**	✗	–	–
DEMIAQYTSALLAG	14	**0.1366**	✗	–	–	✗	–	–	✗	–	–
LQSLQTYVT	9	−**0.0592**	✗	–	–	✗	–	–	✗	–	–
RASANLAATKMSECVLGQ	18	**0.4001**	✗	–	–	✗	–	–	✗	–	–
TDNTFVSGNCD	11	0.0820	✓	14	**0.1793**	✓	14	**0.1793**	✓	14	**0.1793**
KNHTSPDV	8	**0.9006**	✓	8	**0.9006**	✓	8	**0.9006**	✓	8	**0.9006**
GINASVVNIQ	10	**1.0425**	✗	–	–	✗	–	–	✗	–	–
EVAKNLNESL	10	−0.0432	✓	10	−0.0432	✓	14	**0.1512**	✗	–	–
Average		0.4514			0.4762			0.4608			0.6009
Predicted Epitopes in [39]			4V			5V			6V		
EVRQIAPGQTGKIADY	16	**1.3837**	✓	16	**1.3837**	✓	17	1.0936	✓	19	1.1515
TVEKGIYQTSNFRVQP	16	**0.6733**	✓	16	**0.6733**	✓	16	**0.6733**	✗	–	–
HRSYLTPGDSSSGWTA	16	**0.6017**	✓	16	**0.6017**	✓	17	0.4892	✓	17	0.4892
YVGYLQPRTFLLKYNE	16	**0.5108**	✓	18	0.4816	✓	18	0.4816	✓	18	0.4816
CGPKKSTNLVKNKCVN	16	0.2006	✓	20	**0.8935**	✓	20	**0.8935**	✗	–	–
TKTSVDCTMYICGDST	16	0.0937	✓	18	**0.1426**	✓	18	**0.1426**	✗	–	–
TEIYQAGSTPCNGVEG	16	−0.0105	✓	16	−0.0105	✓	16	−0.0105	✓	18	**0.0583**
FERDISTEIYQAGSTP	16	−0.2904	✓	17	−0.1383	✓	17	−0.1383	✓	19	−**0.0782**
FAMQMAYRFNGIGVTQ	16	1.3096	✓	18	**1.4137**	✗	–	–	✗	–	–
IGKIQDSLSSTASALG	16	**0.654**	✓	19	0.5712	✓	19	0.5712	✓	20	0.4992
LQSYGFQPTNGVGYQP	16	**0.5258**	✓	17	0.4203	✓	17	0.4203	✗	–	–
SWMESEFRVYSSANNC	16	**0.1724**	✓	16	**0.1724**	✓	16	**0.1724**	✓	16	**0.1724**
TRFQTLLALHRSYLTP	16	0.5115	✓	18	**0.5595**	✗	–	–	✗	–	–
PQIITTDNTFVSGNCD	16	**0.2404**	✓	16	**0.2404**	✓	16	**0.2404**	✓	16	**0.2404**
QKEIDRLNEVAKNLNE	16	0.0684	✓	18	**0.1255**	✓	18	**0.1255**	✗	–	–
KQIYKTPPIKDFGGFN	16	−**0.2241**	✓	16	−**0.2241**	✓	16	−**0.2241**	✓	16	−**0.2241**
SKRVDFCGK	9	**1.7321**	✓	9	**1.7321**	✓	12	1.3607	✓	12	1.3607
GKYEQY	6	**1.2821**	✓	6	**1.2821**	✓	6	**1.2821**	✓	6	**1.2821**
LDSKVGGNYNYLY	13	**0.8331**	✓	14	0.8329	✓	14	0.8329	✓	14	0.8329
TPGDSSSGWTAGA	13	0.1212	✓	18	**0.5144**	✓	18	**0.5144**	✓	18	**0.5144**
FLPFQ	5	NA	✓	8	**1.4427**	✓	8	**1.4427**	✓	9	1.1432
TSNFRVQPTE	10	**1.3571**	✓	11	1.2323	✓	11	1.2323	✓	11	1.2323
TNLCPF	6	**1.2508**	✓	8	0.8906	✓	13	1.04	✗	–	–
DPSKPSKRSF	10	**0.8148**	✓	10	**0.8148**	✓	10	**0.8148**	✓	11	0.6286
EVFNATRFASVYAWNRKRI	19	**0.2655**	✓	19	**0.2655**	✗	–	–	✗	–	–
AEVQIDR	7	−0.4355	✓	8	−0.2814	✓	11	−**0.0004**	✓	11	−**0.0004**
PTNGVG	6	−1.1441	✓	7	−0.7278	✓	8	−**0.3112**	✓	8	−**0.3112**
QLTPTWRVYSTGSNVFQTRA	20	**0.7725**	✓	20	**0.7725**	✓	20	**0.7725**	✗	–	–
TMSLGAENSVAYSNNS	16	**0.6687**	✓	16	**0.6687**	✓	16	**0.6687**	✓	16	**0.6687**
GFNCYFPLQSY	11	**0.9224**	✓	18	0.8567	✓	18	0.8567	✓	18	0.8567
EPQIITTDNT	10	**0.7545**	✓	13	0.6684	✓	16	0.5227	✓	17	0.3342
NSYECDIPIG	10	0.6533	✓	11	0.8366	✓	11	0.8366	✓	14	**0.9296**
IYKTPPIKDFGGFNF	15	**0.0696**	✓	15	**0.0696**	✓	15	**0.0696**	✓	15	**0.0696**
Average		0.5106			0.5763			**0.6114**			0.5362
Predicted Epitopes in [44]			4V			5V			6V		
LTPGDSSSGWTAG	13	**0.4950**	✓	18	0.3768	✓	18	0.3768	✓	18	0.3768
VRQIAPGQTGKIAD	14	1.2606	✓	15	**1.3487**	✓	16	1.0388	✓	19	1.1515
YQAGSTPCNGV	11	0.0881	✓	13	0.1909	✓	15	**0.2479**	✓	15	**0.2479**
ILPDPSKPSKRS	12	**0.5322**	✓	12	**0.5322**	✓	12	**0.5322**	✓	12	**0.5322**
Average		0.594			**0.6121**			0.5489			0.5771
Predicted Epitopes in [43]			4V			5V			6V		
DVVNQNAQALNTLVKQL	17	**0.0320**	✗	–	–	✗	–	–	✗	–	–
EAEVQIDRLITGRLQSL	17	−0.1784	✓	20	−**0.0881**	✗	–	–	✗	–	–
GAGICASY	8	0.5210	✓	13	**0.6871**	✓	13	**0.6871**	✓	17	0.4587
GSFCTQLN	8	0.8144	✓	9	**0.9306**	✓	9	**0.9306**	✗	–	–

(*continued on next page*)

Table 7 (*continued*).

KGIYQTSN	8	0.2441	✓	8	0.2441	✓	9	**0.4627**	✓	10	0.3992
AMQMAYRF	8	0.9776	✓	9	1.0278	✓	11	**1.2909**	✓	11	**1.2909**
KNHTSPDVDLGDISGIN	17	**1.1116**	✓	18	1.0631	✓	19	0.8800	✓	19	0.8800
AATKMSECVLGQSKRVD	17	**0.6159**	✗	–	–	✗	–	–	✗	–	–
PFAMQMAYRFNGIGVTQ	17	1.3306	✓	18	**1.4137**	✗	–	–	✗	–	–
QALNTLVKQLSSNFGAI	17	**0.0872**	✗	–	–	✗	–	–	✗	–	–
QLIRAAEIRASANLAAT	17	**0.3714**	✓	19	0.3381	✗	–	–	✗	–	–
QQFGRD	6	**−0.5500**	✓	6	**−0.5500**	✓	6	**−0.5500**	✓	6	**−0.5500**
RASANLAATKMSECVLG	17	**0.4414**	✗	–	–	✗	–	–	✗	–	–
RLITGRLQSLQTYVTQQ	17	**−0.2774**	✗	–	–	✗	–	–	✗	–	–
SLQTYVTQQLIRAAEIR	17	**−0.0120**	✗	–	–	✗	–	–	✗	–	–
Average		0.5158			0.5629			**0.6169**			0.4958
Predicted Epitopes in [33]			4V			5V			6V		
DPFLGVYYHKNNKSWME	17	**0.5821**	✓	17	**0.5821**	✓	17	**0.5821**	✓	17	**0.5821**
MDLEGKQGNFKNL	13	**1.2592**	✓	13	**1.2592**	✓	13	**1.2592**	✗	–	–
KHTPINLVRDLPQGFS	16	**0.6403**	✓	17	0.5695	✓	17	0.5695	✗	–	–
TPGDSSSGWTA	11	0.2473	✓	12	0.0746	✓	17	**0.4892**	✓	17	**0.4892**
KSFTVEKGIYQTSNFRVQP	19	**0.5729**	✓	19	**0.5729**	✓	19	**0.5729**	✗	–	–
SNKKFLPF	8	**1.3952**	✓	8	**1.3952**	✓	8	**1.3952**	✓	9	1.1432
TNTSN	5	NA	✓	5		✓	5		✓	5	
NCTEVPVAIHADQLTPT	17	**0.3987**	✗	–	–	✗	–	–	✗	–	–
RVYSTGSNVFQ	11	−0.1000	✓	13	**0.3359**	✓	13	**0.3359**	✗	–	–
VNNSYECDIPI	11	0.6124	✓	16	**0.9123**	✓	16	**0.9123**	✗	–	–
YTMSLGAENSVAYSNN	16	**0.6434**	✓	16	**0.6434**	✓	16	**0.6434**	✓	16	**0.6434**
EQDKNTQ	7	**0.1017**	✓	7	**0.1017**	✓	7	**0.1017**	✓	7	**0.1017**
KQIYKTPPIKDFGGF	15	**−0.3896**	✓	15	**−0.3896**	✓	15	**−0.3896**	✓	15	**−0.3896**
PDPSKPSK	8	**0.0621**	✓	8	**0.0621**	✓	8	**0.0621**	✓	8	**0.0621**
LADAGFIKQYGDCLG	15	**0.2071**	✗	–	–	✗	–	–	✗	–	–
EAEVQ	5	NA	✓	5	NA	✓	5	NA	✓	11	−0.0004
GQSKRVDFC	9	**1.7790**	✓	11	1.4088	✓	12	1.3607	✓	12	1.3607
RNFYEPQIITTD	12	0.3529	✓	15	0.6381	✓	16	**0.6504**	✓	20	0.2624
Average		0.5228			0.5833			**0.6103**			0.4255
Predicted Epitopes in [47]			4V			5V			6V		
RGVYYPDK	8	**1.0191**	✓	8	**1.0191**	✓	8	**1.0191**	✓	11	0.5200
RSSVLHST	8	**0.5459**	✓	10	0.5404	✓	10	0.5404	✗	–	–
DLFLPFFS	8	**−0.3099**	✗	–	–	✗	–	–	✗	–	–
FHAIHV	6	**1.6766**	✓	18	0.6317	✓	18	0.6317	✓	18	0.6317
NPVLPFN	7	**0.5863**	✓	9	0.0146	✓	9	0.0146	✓	9	0.0146
QSLLIVN	7	**0.8168**	✓	15	0.5156	✓	15	0.5156	✗	–	–
NVVIKVCEFQ	10	**−0.1498**	✗	–	–	✗	–	–	✗	–	–
CNDPFLGVYYH	11	0.4109	✓	17	**0.5314**	✓	17	**0.5314**	✓	17	**0.5314**
FEYVSQP	7	**0.9073**	✓	11	0.1016	✓	11	0.1016	✗	–	–
INLVRDL	7	−0.3198	✓	14	0.4022	✓	14	0.4022	✓	17	**0.4924**
LEPLVDLP	8	**−0.3271**	✗	–	–	✗	–	–	✗	–	–
QTLLALHRSY	10	0.5596	✓	17	**0.5921**	✗	–	–	✗	–	–
AAYYVGYL	8	0.5218	✓	12	**0.9255**	✗	–	–	✗	–	–
PRTFLLK	7	−1.3917	✓	10	−0.2800	✓	10	−0.2800	✓	12	**−0.2227**
AVDCALDP	8	**0.7730**	✓	16	0.5804	✓	16	0.5804	✓	16	0.5804
TNLCPFG	7	**1.1812**	✓	8	0.8906	✓	13	1.0400	✗	–	–
SNCVADYSVLYNS	13	−0.1828	✓	13	**0.0152**	✗	–	–	✗	–	–
TFKCYGVSPT	10	**1.5059**	✓	20	0.8913	✓	20	0.8913	✗	–	–
TGCVIA	6	**0.4716**	✓	10	0.0996	✓	13	−0.1592	✓	14	−0.1234
CYFPLQSY	8	**0.9394**	✓	8	**0.9394**	✓	12	0.8719	✓	12	0.8719
FGGVSVIT	8	**0.7715**	✓	12	0.4578	✓	13	0.4931	✓	13	0.4931
CTEVPVAIHAD	11	**0.0499**	✗	–	–	✗	–	–	✗	–	–
AGCLIGA	7	**0.1743**	✗	–	–	✗	–	–	✗	–	–
GAGICASY	8	0.5210	✓	13	**0.6871**	✓	13	**0.6871**	✓	17	0.4587
VASQSII	7	−0.0188	✓	16	0.3257	✓	16	0.3257	✓	18	**0.4018**
TTEILPVS	8	**1.2071**	✗	–	–	✗	–	–	✗	–	–
SVDCTMY	7	**1.0932**	✓	17	−0.0258	✓	18	0.1426	✗	–	–
SNLLLQYGSFCTQL	14	**0.7599**	✗	–	–	✗	–	–	✗	–	–
VFAQVKQI	8	**0.5854**	✓	14	0.3493	✓	15	0.4451	✗	–	–
SQILPD	6	−0.1542	✓	8	0.0383	✓	8	0.0383	✓	11	**0.5569**
YGDCLGD	7	−0.5555	✓	12	**0.5494**	✓	14	0.0416	✗	–	–
RDLICAQ	7	**1.1443**	✗	–	–	✗	–	–	✗	–	–
LTVLPPL	7	**0.6786**	✗	–	–	✗	–	–	✗	–	–
YTSALLAG	8	**0.3798**	✓	20	0.3640	✗	–	–	✗	–	–
LNTLVKQL	8	−0.7591	✓	16	**−0.0646**	✓	16	**−0.0646**	✗	–	–
ISSVLND	7	0.0414	✓	11	**0.7339**	✓	12	0.6035	✗	–	–
SLQTYVTQQ	9	**−0.0089**	✗	–	–	✗	–	–	✗	–	–
SECVLGQS	8	−0.0110	✓	13	**0.5417**	✗	–	–	✗	–	–
PHGVVFLHVTYVPA	14	**0.8058**	✗	–	–	✗	–	–	✗	–	–

(*continued on next page*)

Table 7 (*continued*).

PAICHDG	7	−1.0100	✓	15	**0.2145**	✓	15	**0.2145**	✗	–	–
SGNCDVVIGI	10	**0.7421**	✗	–	–	✗	–	–	✗	–	–
ASVVNI	6	**0.8671**	✓	13	0.1922	✗	–	–	✗	–	–
Average		0.3938			**0.4258**			0.4011			0.4005
Predicted Epitopes in [45]			4V			5V			6V		
VRQIAPGQTGKIAD	14	1.2606	✓	15	**1.3487**	✓	16	1.0388	✓	19	1.1515
VLGQSKRVDFCGKG	14	**1.3582**	✗	–	–	✗	–	–	✗	–	–
GLTGTGVLTESNKK	14	**1.0227**	✓	14	**1.0227**	✓	14	**1.0227**	✓	16	0.6686
KIADYNYKLPDDFT	14	**0.9567**	✓	14	**0.9567**	✓	14	**0.9567**	✓	14	**0.9567**
Average		**1.1495**			1.1094			1.006			0.9256
Predicted Epitopes in [46]			4V			5V			6V		
DPFLGVYYHKNNKSWME	17	**0.5821**	✓	17	**0.5821**	✓	17	**0.5821**	✓	17	**0.5821**
MDLEGKQGNFKNL	13	**1.2592**	✓	13	**1.2592**	✓	13	**1.2592**	✗	–	–
KHTPINLVRDLPQGFS	16	**0.6403**	✓	17	0.5695	✓	17	0.5695	✗	–	–
TPGDSSSGWTA	11	0.2473	✓	12	0.0746	✓	17	**0.4892**	✓	17	**0.4892**
KSFTVEKGIYQTSNFRVQP	19	**0.5729**	✓	19	**0.5729**	✓	19	**0.5729**	✗	–	–
VNNSYECDIPI	11	0.6124	✓	16	**0.9123**	✓	16	**0.9123**	✗	–	–
YTMSLGAENSVAYSNN	16	**0.6434**	✓	16	**0.6434**	✓	16	**0.6434**	✓	16	**0.6434**
Average		0.6111			0.6591			**0.7184**			0.5716

them. Considering the antigenicity scores of epitopes of different lengths, the 5V method detected more antigen epitopes.

When the results of the methods presented in the literature are compared with the developed method, it is seen that most of the epitopes suggested as vaccine candidates for SARS-CoV-2 with many different methods can also be detected by the ensemble fuzzy method. This shows that the proposed ensemble fuzzy method is robust. Among the 4V, 5V, and 6V methods obtained by combining the decisions with different majority decisions, the 5V method was generally more successful in terms of the average antigenicity score.

5. Discussion

SARS-CoV-2 B-cell epitope identification with the aid of a high-performance prediction method contributes to rapid, reliable, and effective protein-based vaccine development. The use of experimental methods in vaccine development is quite time-consuming, costly, and labor-intensive. Therefore, the main aim of this study was to propose a method that can predict B-cell epitopes with high accuracy. The results obtained from the study show that we have achieved this goal. The SARS-CoV B-cell epitope prediction of the five different fuzzy learning classification methods in different test data minimal error rate was 7.5% (SLAVE), and the maximal error rate was 52.5% (W), while the minimal error rate of the ensemble fuzzy method was 5% and the maximal error rate was 12.5%. When the average errors of the test results were compared, the proposed method had the lowest error rate of 8.33%, followed by the SLAVE method at 20.42%. The mean error of the most successful fuzzy learning model in SARS-CoV B-cell epitope prediction was approximately 2.5 times higher than of the proposed model. From the obtained results, it is clearly seen that the proposed method outperformed other methods. This shows that ensemble learning methods are more successful than individual methods.

The main advantage of the proposed method is that the decisions made by fuzzy methods are combined with an ensemble-based structure. The fuzzy methods used include different learning and decision-making approaches, as explained in Section 3.2. It has been seen in the majority voting and decision aggregation phase that different fuzzy methods learn different features in the dataset. Thus, the average spread of a model that contributes to the ensemble is reduced, and the average prediction performance for each model is improved. In fact, as clearly shown in Table 3 the aggregation of decisions improved the estimation performance of single fuzzy classifiers.

The prediction accuracies of the studies on B-cell epitope prediction in the literature were compared with the prediction accuracies obtained from this study. In [54], different machine learning methods were compared and it was reported that the most successful method was the ensemble method with an accuracy value of 87.8%. In [51], Bayesian neural networks with drop-weight models were proposed for epitope prediction. The prediction accuracy of the proposed model was 85%. In [53], the attentional mechanism LSTM network approach was used for epitope prediction. With this model, epitopes were predicted with 79% accuracy. In this study, the proposed ensemble classifier outperformed other studies in the literature, with a minimum accuracy of 87.5% and maximum accuracy of 95.0% epitope prediction.

In this study, we determined that the dataset was imbalanced; therefore, we applied the subsampling preprocessing step. Thus, we randomly generated 6 different sub-datasets with the positive and negative class labels of the samples balanced. The SARS-CoV B-cell epitope prediction accuracies of the proposed ensemble fuzzy model are illustrated in Fig. 9.

The dataset used in the study consists of all possible subsequences of the SARS-CoV-2 spike protein of different lengths, and there is no label information to definitively determine whether a sequence is an epitope or not. Fuzzy logic approaches enable us to take into account imprecise information while making a decision. As indicated in the predictions, it can measure more sensitively than classical logic-based classification methods. The results obtained showed that it was more successful than other machine learning methods evaluated in the literature.

In addition, the SARS-CoV-2 B-cell epitope prediction results obtained in this study were validated with the results reported in the literature and the epitope results predicted by the BepiPred server. Furthermore, antigenicity scores of SARS-CoV-2 B-cell epitopes were measured. Thus, we hope that the information obtained from this study will help develop an effective protein-based vaccine against SARS-CoV-2.

6. Conclusion

In this study, an efficient fuzzy learning-based model is proposed that predicts potential epitopes to assist the first stage of mRNA-based vaccine development. The SARS-CoV B-cell epitope prediction performances of five different fuzzy learning methods were examined and compared with the proposed method. According to the results obtained, the proposed ensemble fuzzy

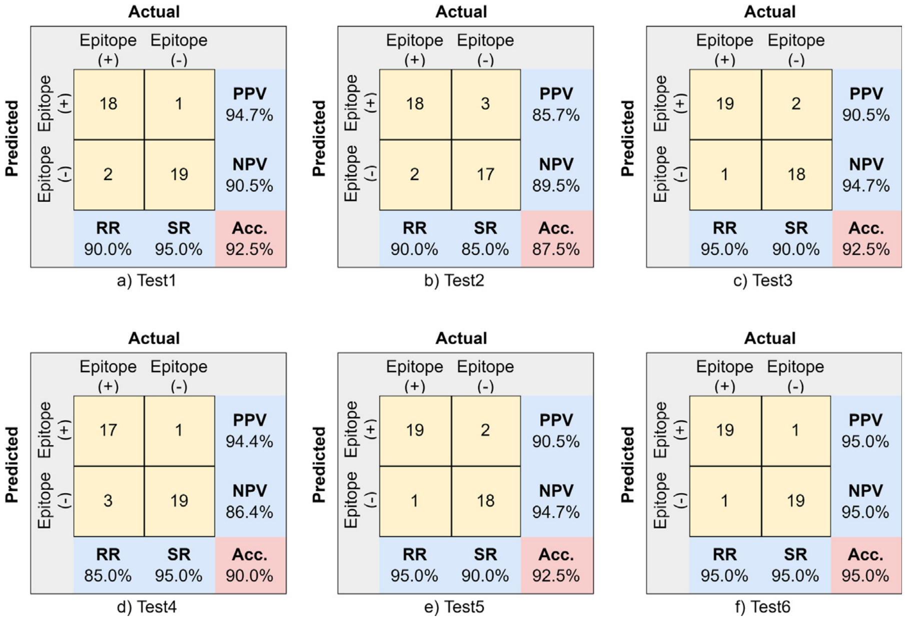

Fig. 9. Confusion matrices of the proposed ensemble fuzzy model for different test sets.

method showed superior performance with maximal 95% accuracy and average 91.7% accuracy to other methods. After that, with the proposed method, B-cell epitope prediction was made in unlabeled SARS-CoV-2 data, and the epitopes found were confirmed with those reported in the literature and the predictions of the BepiPred server. Moreover, antigenicity scores were measured for protein sequences of epitopes identified using the VaxiJen server.

The virus is still spreading rapidly and therefore mutating. It has been determined that some mutations can escape vaccines [15,62]. The fact that the virus mutates and may require the redevelopment of vaccines increases the importance of using in silico methods. However, if these mutations occur outside of the identified epitope regions, the results will not be affected. Therefore, the identified epitopes can be used as potential antigens from which more detailed assays can be conducted in vitro to evaluate vaccine efficacy. It is anticipated that the information obtained from this study will contribute to the development of vaccines against different epidemics that may occur in the future, especially SARS-CoV-2 and its possible mutations.

Code availability statement

The source code has been available at https://github.com/ZBaOz/Epitope-Identification.

CRediT authorship contribution statement

Zeynep Banu Ozger: Conceptualization, Writing – original draft, Methodology, Validation, Software, Visualization, Writing – review & editing. **Pınar Cihan:** Writing – original draft, Methodology, Validation, Software, Visualization, Writing – review & editing.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

This study was supported by The Scientific and Technological Research Council of Turkey-TÜBİTAK (Project Number: 121E326).

Appendix A. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.asoc.2021.108280.

References

[1] C.A. Janeway Jr., P. Travers, M. Walport, M.J. Shlomchik, Principles of innate and adaptive immunity, in: Immunobiology: The Immune System in Health and Disease, fifth ed., Garland Science, 2001.
[2] J.S. Marshall, R. Warrington, W. Watson, H.L. Kim, An introduction to immunology and immunopathology, Allergy Asthma Clin. Immunol. 14 (2) (2018) 1–10.
[3] D.D. Chaplin, Overview of the immune response, J. Allergy Clin. Immunol. 125 (2) (2010) S3–S23.
[4] M.C. Jespersen, B. Peters, M. Nielsen, P. Marcatili, Bepipred-2.0: improving sequence-based b-cell epitope prediction using conformational epitopes, Nucleic Acids Res. 45 (W1) (2017) W24–W29.
[5] W.E. Paul, Fundamental immunology, in: Fundamental Immunology, 1985, p. 809.
[6] J.V. Ponomarenko, M.H. Van Regenmortel, B cell epitope prediction, Struct. Bioinform. 2 (2009) 849–879.
[7] J.L. Sanchez-Trincado, M. Gomez-Perosanz, P.A. Reche, Fundamentals and methods for t-and b-cell epitope prediction, J. Immunol. Res. 2017 (2017).
[8] M. Levitt, Nature of the protein universe, Proc. Natl. Acad. Sci. 106 (27) (2009) 11079–11084.
[9] F. Li, Structure, function, and evolution of coronavirus spike proteins, Annu. Rev. Virol. 3 (2016) 237–261.
[10] E. De Wit, N. Van Doremalen, D. Falzarano, V.J. Munster, Sars and mers: recent insights into emerging coronaviruses, Nat. Rev. Microbiol. 14 (8) (2016) 523.
[11] C.S.G. of the International, et al., The species severe acute respiratory syndrome-related coronavirus: classifying 2019-ncov and naming it sars-cov-2, Nat. Microbiol. 5 (4) (2020) 536.
[12] M. Galanopoulos, A. Doukatas, M. Gazouli, Origin and genomic characteristics of sars-cov-2 and its interaction with angiotensin converting enzyme type 2 receptors, focusing on the gastrointestinal tract, World J. Gastroenterol. 26 (41) (2020) 6335.
[13] M.T. Ul Qamar, S. Saleem, U.A. Ashfaq, A. Bari, F. Anwar, S. Alqahtani, Epitope-based peptide vaccine design and target site depiction against middle east respiratory syndrome coronavirus: an immune-informatics study, J. Transl. Med. 17 (1) (2019) 1–14.

[14] P. Ellis, F. Somogyvári, D.P. Virok, M. Noseda, G.R. McLean, Decoding covid-19 with the sars-cov-2 genome, Curr. Genet. Med. Rep. (2021) 1–12.

[15] P. Cihan, Forecasting fully vaccinated people against COVID-19 and examining future vaccination rate for herd immunity in the US, Asia, europe, africa, south america, and the world, Appl. Soft Comput. 111 (2021) 107708.

[16] C. Chakraborty, A. Sharma, G. Sharma, M. Bhattacharya, S. Lee, Sars-CoV-2 causing pneumonia-associated respiratory disorder (COVID-19): diagnostic and proposed therapeutic options, Eur. Rev. Med. Pharmacol. Sci. 24 (7) (2020) 4016–4026.

[17] A. Grifoni, J. Sidney, Y. Zhang, R.H. Scheuermann, B. Peters, A. Sette, A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2, Cell Host Microbe 27 (4) (2020) 671–680.

[18] S.C. Jordan, Innate and adaptive immune responses to SARS-CoV-2 in humans: relevance to acquired immunity and vaccine responses, Clin. Exp. Immunol. 204 (3) (2021) 310–320.

[19] C.O. Barnes, C.A. Jette, M.E. Abernathy, K.-M.A. Dam, S.R. Esswein, H.B. Gristick, A.G. Malyutin, N.G. Sharaf, K.E. Huey-Tubman, Y.E. Lee, et al., Sars-CoV-2 neutralizing antibody structures inform therapeutic strategies, Nature 588 (7839) (2020) 682–687.

[20] J.M. Dan, J. Mateus, Y. Kato, K.M. Hastie, E.D. Yu, C.E. Faliti, A. Grifoni, S.I. Ramirez, S. Haupt, A. Frazier, et al., Immunological memory to SARS-CoV-2 assessed for up to 8 months after infection, Science 371 (6529) (2021).

[21] E. Ong, M.U. Wong, A. Huffman, Y. He, Covid-19 coronavirus vaccine design using reverse vaccinology and machine learning, Front. Immunol. 11 (2020) 1581.

[22] Y. Zhou, S. Jiang, L. Du, Prospects for a mers-cov spike vaccine, Expert Rev. Vaccines 17 (8) (2018) 677–686.

[23] L. Du, G. Zhao, Y. Lin, H. Sui, C. Chan, S. Ma, Y. He, S. Jiang, C. Wu, K.-Y. Yuen, et al., Intranasal vaccination of recombinant adeno-associated virus encoding receptor-binding domain of severe acute respiratory syndrome coronavirus (sars-cov) spike protein induces strong mucosal immune responses and provides long-term protection against sars-cov infection, J. Immunol. 180 (2) (2008) 948–956.

[24] C.H. Lee, H. Koohy, In silico identification of vaccine targets for 2019-ncov, F1000Research 9 (2020).

[25] Z. Ceylan, Estimation of covid-19 prevalence in italy, spain, and france, Sci. Total Environ. 729 (2020) 138817.

[26] Z. Malki, E.-S. Atlam, A. Ewis, G. Dagnew, A.R. Alzighaibi, G. ELmarhomy, M.A. Elhosseini, A.E. Hassanien, I. Gad, Arima models for predicting the end of covid-19 pandemic and the risk of second rebound, Neural Comput. Appl. 33 (7) (2021) 2929–2948.

[27] V.K.R. Chimmula, L. Zhang, Time series forecasting of covid-19 transmission in canada using lstm networks, Chaos Solitons Fractals 135 (2020) 109864.

[28] P. Cihan, Fuzzy rule-based system for predicting daily case in covid-19 outbreak, in: 2020 4th International Symposium on Multidisciplinary Studies and Innovative Technologies (ISMSIT), IEEE, 2020, pp. 1–4.

[29] F. Demir, Deepcoronet: A deep lstm approach for automated detection of covid-19 cases from chest x-ray images, Appl. Soft Comput. 103 (2021) 107160.

[30] A. Saygılı, A new approach for computer-aided detection of coronavirus (covid-19) from ct and x-ray images using machine learning methods, Appl. Soft Comput. 105 (2021) 107323.

[31] A. Castiglione, P. Vijayakumar, M. Nappi, S. Sadiq, M. Umer, Covid-19: Automatic detection of the novel coronavirus disease from ct images using an optimized convolutional neural network, IEEE Trans. Ind. Inf. (2021).

[32] M.S. Sohail, S.F. Ahmed, A.A. Quadeer, M.R. McKay, In silico t cell epitope identification for sars-cov-2: Progress and perspectives, Adv. Drug Deliv. Rev. (2021).

[33] M. Bhattacharya, A.R. Sharma, P. Patra, P. Ghosh, G. Sharma, B.C. Patra, S.-S. Lee, C. Chakraborty, Development of epitope-based peptide vaccine against novel coronavirus 2019 (sars-cov-2): Immunoinformatics approach, J. Med. Virol. 92 (6) (2020) 618–631.

[34] M. Surjit, S.K. Lal, The sars-cov nucleocapsid protein: a protein with multifarious activities, Infect. Genet. Evol. 8 (4) (2008) 397–405.

[35] U.J. Buchholz, A. Bukreyev, L. Yang, E.W. Lamirande, B.R. Murphy, K. Subbarao, P.L. Collins, Contributions of the structural proteins of severe acute respiratory syndrome coronavirus to protective immunity, Proc. Natl. Acad. Sci. 101 (26) (2004) 9804–9809.

[36] R. Vita, S. Mahajan, J.A. Overton, S.K. Dhanda, S. Martini, J.R. Cantrell, D.K. Wheeler, A. Sette, B. Peters, The immune epitope database (iedb): 2018 update, Nucleic Acids Res. 47 (D1) (2019) D339–D343.

[37] B.E. Pickett, E.L. Sadat, Y. Zhang, J.M. Noronha, R.B. Squires, V. Hunt, M. Liu, S. Kumar, S. Zaremba, Z. Gu, et al., Vipr: an open bioinformatics database and analysis resource for virology research, Nucleic Acids Res. 40 (D1) (2012) D593–D598.

[38] J.V. Kringelum, C. Lundegaard, O. Lund, M. Nielsen, Reliable b cell epitope predictions: impacts of method development and improved benchmarking, PLoS Comput. Biol. 8 (12) (2012) e1002829.

[39] H.-Z. Chen, L.-L. Tang, X.-L. Yu, J. Zhou, Y.-F. Chang, X. Wu, Bioinformatics analysis of epitope-based vaccine design against the novel sars-cov-2, Infect. Dis. Poverty 9 (1) (2020) 1–10.

[40] S. Saha, G.P.S. Raghava, Prediction of continuous b-cell epitopes in an antigen using recurrent neural network, Proteins: Struct. Funct. Bioinform. 65 (1) (2006) 40–48.

[41] I.A. Doytchinova, D.R. Flower, Vaxijen: a server for prediction of protective antigens, tumour antigens and subunit vaccines, BMC Bioinformatics 8 (1) (2007) 1–7.

[42] V. Baruah, S. Bose, Immunoinformatics-aided identification of t cell and b cell epitopes in the surface glycoprotein of 2019-nCoV, J. Med. Virol. 92 (5) (2020) 495–500.

[43] S.F. Ahmed, A.A. Quadeer, M.R. McKay, Preliminary identification of potential vaccine targets for the covid-19 coronavirus (sars-cov-2) based on sars-cov immunological studies, Viruses 12 (3) (2020) 254.

[44] B. Sarkar, M.A. Ullah, F.T. Johora, M.A. Taniya, Y. Araf, The essential facts of wuhan novel coronavirus outbreak in china and epitope-based vaccine designing against 2019-ncov, BioRxiv (2020).

[45] L. Lin, S. Ting, H. Yufei, L. Wendong, F. Yubo, Z. Jing, Epitope-based peptide vaccines predicted against novel coronavirus disease caused by sars-cov-2, Virus Res. 288 (2020) 198082.

[46] S. Ismail, S. Ahmad, S.S. Azam, Immunoinformatics characterization of sars-cov-2 spike glycoprotein for prioritization of epitope based multivalent peptide vaccine, J. Molecular Liquids 314 (2020) 113612.

[47] H.M. Rehman, M.U. Mirza, M.A. Ahmad, M. Saleem, M. Froeyen, S. Ahmad, R. Gul, H.A. Alghamdi, M.S. Aslam, M. Sajjad, et al., A putative prophylactic solution for covid-19: Development of novel multiepitope vaccine candidate against sars-cov-2 by comprehensive immunoinformatic and molecular modelling approach, Biology 9 (9) (2020) 296.

[48] M.S. Shoukat, A.D. Foers, S. Woodmansey, S.C. Evans, A. Fowler, E.J. Soilleux, Use of machine learning to identify a t cell response to sars-cov-2, Cell Rep. Med. 2 (2) (2021) 100192.

[49] V. Jurtz, S. Paul, M. Andreatta, P. Marcatili, B. Peters, M. Nielsen, Netmhcpan-4.0: improved peptide–mhc class i interaction predictions integrating eluted ligand and peptide binding affinity data, J. Immunol. 199 (9) (2017) 3360–3368.

[50] M.V. Pogorelyy, A.D. Fedorova, J.E. McLaren, K. Ladell, D.V. Bagaev, A.V. Eliseev, A.I. Mikelov, A.E. Koneva, I.V. Zvyagin, D.A. Price, et al., Exploring the pre-immune landscape of antigen-specific t cells, Genome Med. 10 (1) (2018) 1–14.

[51] B. Ghoshal, B. Ghoshal, S. Swift, A. Tucker, Uncertainty estimation in sars-cov-2 b-cell epitope prediction for vaccine development, 2021, arXiv preprint arXiv:2103.11214.

[52] F. Corporation, Covid-19/sars b-cell epitope prediction, 0000. URL https://www.kaggle.com/futurecorporation/epitope-prediction.

[53] T. Noumi, S. Inoue, H. Fujita, K. Sadamitsu, M. Sakaguchi, A. Tenma, H. Nakagami, Epitope prediction of antigen protein using attention-based lstm network, J. Inf. Process. 29 (2021) 321–327.

[54] N. Jain, S. Jhunthra, H. Garg, V. Gupta, S. Mohan, A. Ahmadian, S. Salahshour, M. Ferrara, Prediction modelling of covid using machine learning methods from b-cell dataset, Results Phys. 21 (2021) 103813.

[55] F. Krammer, Sars-cov-2 vaccines in development, Nature 586 (7830) (2020) 516–527.

[56] H. Ishibuchi, T. Nakashima, T. Murata, Performance evaluation of fuzzy classifier systems for multidimensional pattern classification problems, IEEE Trans. Syst. Man Cybern. B 29 (5) (1999) 601–618.

[57] H. Ishibuchi, T. Yamamoto, T. Nakashima, Hybridization of fuzzy gbml approaches for pattern classification problems, IEEE Trans. Syst. Man Cybern. B 35 (2) (2005) 359–365.

[58] Z. Chi, H. Yan, T. Pham, Fuzzy Algorithms: With Applications to Image Processing and Pattern Recognition, Vol. 10, World Scientific, 1996.

[59] L.-X. Wang, J.M. Mendel, Generating fuzzy rules by learning from examples, IEEE Trans. Syst. Man Cybern. 22 (6) (1992) 1414–1427.

[60] H. Ishibuchi, T. Nakashima, Effect of rule weights in fuzzy rule-based classification systems, IEEE Trans. Fuzzy Syst. 9 (4) (2001) 506–515.

[61] A. González, R. Pérez, Selection of relevant features in a fuzzy genetic learning algorithm, IEEE Trans. Syst. Man Cybern. B 31 (3) (2001) 417–425.

[62] E.C. Thomson, L.E. Rosen, J.G. Shepherd, R. Spreafico, A. da Silva Filipe, J.A. Wojcechowskyj, C. Davis, L. Piccoli, D.J. Pascall, J. Dillen, et al., Circulating sars-cov-2 spike n439k variants maintain fitness while evading antibody-mediated immunity, Cell 184 (5) (2021) 1171–1187.