Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Prediction of viral-host interactions of COVID-19 by computational methods

Talha Burak Alakus a, *, Ibrahim Turkoglu b

a Kirklareli University, Department of Software Engineering, Kirklareli, 39000, Turkey
b Firat University, Department of Software Engineering, Elazig, 23119, Turkey

ARTICLE INFO

Keywords:
Protein mapping
Deep learning
Covid-19
SARS-CoV-2 virus

ABSTRACT

Experimental approaches are currently used to determine viral-host interactions, but these approaches are both time-consuming and costly. For these reasons, computational-based approaches are recommended. In this study, using computational-based approaches, viral-host interactions of SARS-CoV-2 virus and human proteins were predicted. The study consists of four different stages; in the first stage viral and host protein sequences were obtained. In the second stage, protein sequences were converted into numerical expressions by various protein mapping methods. These methods are entropy-based, AVL-tree, FIBHASH, binary encoding, CPNR, PAM250, BLOSUM62, Atchley factors, Meller parameters, IIIP, AESNN1, Miyazawa energies, Micheletti potentials, Z-scale, and hydrophobicity. In the third stage, a deep learning model was designed and BiLSTM was used for this. In the last stage, the protein sequences were classified, and the viral-host interactions were predicted. The performances of protein mapping methods were determined by accuracy, F1-score, specificity, sensitivity, and AUC scores. According to the classification results, the best classification process was obtained by the entropy-based method. With this method, 94.74% accuracy, and 0.95 AUC score were calculated. Then, the most successful classification process was performed with the Z-scale and 91.23% accuracy, and 0.96 AUC score were obtained. Although other protein mapping methods are not as efficient as Z-scale and entropy-based methods, they have achieved successful classification. AVL-tree, FIBHASH, binary encoding, CPNR, PAM250, BLOSUM62, Atchley factors, Meller parameters and AESNN1 methods showed over 80% accuracy, F1-score, and AUC score. Accuracy scores of EIIP, Miyazawa energies, Micheletti potentials and hydrophobicity methods remained below 80%. When the results were examined in general, it was observed that the computational approaches were successful in predicting viral-host interactions between SARS-CoV-2 virus and human proteins.

1. Introduction

The SARS-CoV-2 virus, which emerged in 2019 in Wuhan, China, spread all over the world in a short time and became a pandemic by causing COVID-19 disease. This disease, which affects the immune system, causes fever and cough in people, and in some serious cases, it causes pneumonia [1]. Although it is a serious disease, approximately 80% of people have mild to moderate disease and do not need any treatment [2,3]. However, people whose immune system is not strong enough and who have chronic disorders have this disease severely and in very serious cases, they are hospitalized in intensive care units [4]. COVID-19 disease is considered a serious illness because it is highly permeable and contagious. In addition, the COVID-19 disease is seriously affecting health systems. Furthermore, the scarcity of intensive care units, long treatment periods and insufficient hospital resources clearly demonstrate the impact of this epidemic on the health system [5]. Today, vaccines have been developed and started to be used to combat COVID-19. However, although vaccines are developed, the virus rebuilds itself and constantly reveals various variants. Some studies have shown that the efficacy of vaccines and antibodies against new variants is reduced [6,7]. It has also been observed that the new variants are more contagious [8]. For these reasons, it is important to diagnose the disease at an early stage to prevent such problems. To fight COVID-19 disease, the genetic characteristics of SARS-CoV-2 virus should be well known. Thanks to the genetic characteristics of the virus, various drug and vaccine treatments related to COVID-19 disease can be revealed [9,10]. SARS-CoV-2 virus consists of four main structural proteins; spike glycoprotein (S), membrane glycoprotein (M), small...
envelope glycoprotein (E), and nucleocapsid protein (N) [11]. In Fig. 1, the genetic structure of the SARS-CoV-2 virus is given.

Understanding how these virus proteins interact with host cells for survival and reproduction is of great importance for the development and use of drugs [13]. Viruses communicate with host cells through protein-protein interactions. Protein interactions between viral and host are responsible for the entire life cycle of the virus [14]. This cycle includes infection of the host cell and replication of the viral genome. Therefore, the identification of protein interactions between virus and host proteins provides information about how virus proteins work, how they replicate, and how they cause disease. Analysis of interactions between proteins elucidates processes such as immunity, cellular signaling, replication, and cellular division. In addition, viral-host interactions constitute a potential target for antiviral therapies [15]. Experimental approaches are frequently used to determine interactions between proteins and between virus and host. These approaches generate a large amount of data and are costly in terms of both time and laboratory equipment [16]. The most used experimental methods are YTH (Yeast Two Hybrid), GST Pull-Down, NAPPA (Nucleic Acid Programmable Protein Array) and APMS (Affinity Purification Mass Spectrometry) [14]. Yet, protein interactions vary according to the experimental methods. Because experimental methods are based on both environmental and operational processes, false positive and false negative results may occur [17]. Because of these problems and disadvantages, computational-based approaches are more preferred, and their popularity is constantly increasing [18,19].

Computational approaches were also used in this study and viral-host interactions between SARS-CoV-2 and human proteins were predicted. The study consists of four stages. In the first stage, viral-host interaction data between SARS-CoV-2 and human were obtained. In the second stage, the protein sequences were mapped by numerical mapping methods. For this purpose, fifteen protein mapping methods were used. These methods are AESNN1, Atchley factors, AVL-tree (Adelson, Velsky and Landis), binary encoding, BLOSUM62 (Blocks Substitution Matrix), CPNR (Complex Prime Number Representation), EIIP (Electron-Ion Interaction Potential), FIBHASH (FibonacciHash), hydrophobicity, entropy-based, Meiler parameters, Micheletti potentials, Miyazawa energies, PAM250 (Point Accepted Mutation), and Z-scale. In the third stage, a deep learning model was designed and viral-host interactions between SARS-CoV-2 virus and human proteins were classified. In the last stage, the performances of the numerical mapping methods were determined by the evaluation criteria of accuracy, sensitivity, specificity, AUC score and F1-score. The highlights of the study can be summarized as follows:

- In this study, viral-host interactions between SARS-CoV-2 and human proteins were predicted by computational approaches.
- To the best of our knowledge, proteins were analyzed for the first time in this study using AVL-tree, entropy-based, and FIBHASH methods.
- In this study, viral-host interactions between SARS-CoV-2 and human proteins were effectively classified by deep learning.

The remainder of the study is organized as follows. In the second section, the studies carried out in this field in the literature are examined and information about the methods and classifiers is given. In the third section, information about the protein mapping methods is given. In this section, the data set is also emphasized, and necessary explanations are made about the training and test data. In this section, the designed deep learning model is also examined, and its parameters are explained. In the fourth section, the classification process was carried out and the performances of the protein mapping methods were compared by giving the classification results. In the last section information is given about the contribution of the study to the literature and the fields of study that can be used in the future.

2. Related works

In this section, studies predicting viral-host interactions between SARS-CoV-2 and human proteins are reviewed. The data, methods and classification results used in these studies were analyzed. In study [13], researchers predicted viral-host interactions between SARS-CoV-2 and human proteins using machine learning techniques. In the study, SVM (Support Vector Machine), KNN (K Nearest Neighbor), NB (Naïve Bayes), and RF (Random Forest) algorithms were used for classification. Protein sequences were mapped by AAC (Amino Acid Composition), CT (Conjoint Triad) and PseAAC (Pseudo Amino Acid Composition) methods and features were obtained. Then, feature selection was performed with the LVQ (Learning Vector Quantization) method. The performances of the classification algorithms were measured with accuracy, recall, specificity, precision, and F1-score. The best results were obtained with the ensemble model which consists of SVM and RF algorithms. With the ensemble model, an accuracy of 72.33% was obtained. Researchers in study [20] analyzed the host-cell interactome of SARS-CoV-2 infection. In the study, a network-based approach was used for the analysis process. Data were obtained from the Reactome and KEGG datasets. The pathogen of SARS-CoV-2 virus was predicted using three human coronavirus molecules (SARS-CoV, MERS-COV and HCoV-229E) and protein-protein interactions. At the end of the study, it was stated that the developed method was successful and could be used to identify potential new biological targets. In another study, researchers aligned viral-host protein-protein interactions of SARS-CoV-1, SARS-CoV-2 viruses and human proteins using integer linear programming [21]. In the study, gene ontologies were used and protein interactions were examined in terms of MF (Molecular Function), CC (Cellular Component) and BP (Biological Process). In the first stage of the study, the proteins of SARS-CoV-1 and SARS-CoV-2 viruses were aligned. Alignment was performed separately for each gene ontology. After the alignment process, human proteins were used and the proteins with which these viruses interacted were determined. The best BP and CC results for spike protein (S) strains were obtained with Q7Z5G4 protein. For MF, it was obtained from Q7Z5G4 and Q9C085 proteins. For the Envelope protein (E) strain, the best alignment result was obtained with the O00203 protein in the BB category. While the Q8IWAS protein gave the best alignment result in the CC category, the best alignment interaction was calculated for the O00203 protein in the MF category. In the membrane protein (M), the best alignment score was obtained from the Q96DS3 protein in the CC category. In the BP category, the best score was obtained from the P27105 protein, while in the MF category it was again obtained from the Q96DS3 protein. In the nucleocapsid protein (N) type, the best alignment score was obtained from the P67870 protein.
for BP. The most efficient alignment process in the CC gene ontology type was calculated with the P11940 protein. It was observed that multiple proteins from the MF category were effective in the alignment process. At the end of the study, it was stated that the alignment process was effective in determining virus-host interactions of SARS-CoV-2 and human proteins. In study [22], researchers developed a novel deep learning model and predicted virus-host interactions from protein sequences and infectious disease phenotypes. In the study, human proteins and virus taxonomies were mapped by one-hot encoding method. After the mapping process, a CNN (Convolutional Neural Networks) based deep learning model was developed and the data were classified. The classification process was carried out in three different categories as family-based, taxon-based, and protein-based, and the performance of the proposed deep learning model was calculated with the ROC (Receiver Characteristic Curve). While 81.30% success was achieved in family-based classification, this rate was calculated as 82.90% in taxon-based classification. In protein-based classification, this rate was calculated as 80%. At the end of the study, it was observed that the proposed deep learning model effectively determined the viral-host interactions between SARS-CoV-2 and human proteins. In another study, researchers predicted viral-host interactions between SARS-CoV-2 virus and human proteins using a network-based method [23]. Structural and content information of virus and human proteins were used, and this information was combined. Then, classification was performed with MLP (Multi-Layer Perceptron) and the model was compared with various models in the literature. The performance of the classification algorithm was calculated with accuracy, precision, recall, F1-score, and AUC values. With the proposed deep learning model, an accuracy score of 97.1% was obtained, while an AUC score of 0.997 was calculated.

3. Materials and methods

3.1. SARS-CoV-2 and human protein-protein interaction data

In this study, the data set given in the study [24] was used. In the study, the researchers shared a data set containing protein-protein interactions between the SARS-CoV-2 virus and human proteins. Affinity purification mass spectrometry was used to create this data set. In the data set, there are 332 interactions between 332 human proteins and 4 structural and 20 coronavirus proteins. This data set was used to construct positive training and test data sets in the study. There is no specific gold standard for constructing the negative data set in protein-protein interactions. Generally, two methods, random matching, and sub-cellular localization, are used when creating a negative data set [13]. In most studies, non-interacting proteins are randomly generated, and protein pairs used in the positive dataset are eliminated [25,26]. On the other hand, in some studies, negative data set is created by sub-cellular localization [27]. However, negative data sets generated by these two methods are not very reliable. In the random matching process, positive interactions can be mistakenly treated as negative interactions, and different accuracy scores can be calculated for different random matching processes. In addition, biased accuracy scores can be obtained with sub-cellular localization [26]. For such reasons, these two methods were not used while creating the negative data set in this study. The negative data set was obtained from study [13]. In the study [13], the researchers used a degree-based approach when constructing the negative data set. It has been determined that the negative data set created with this approach is more effective both in the training phase and in the testing phase compared to both random matching and sub-cellular localization methods. While a total of 1,327 protein sequences were used in the positive data set, 4,554 protein sequences were used for the negative data set.

3.2. Protein mapping methods

In this study, protein sequences were mapped various protein mapping methods. These methods are AESNN1, Atchley factors, hydrophobicity, Meiler parameters, EIIIP, CPNR, binary-encoding, FIB-HASH, PAM250, BLOSUM62, Michelei potentials, Miyazawa energies, entropy-based, AVL-tree, and Z-scale. In the following sections, information about these methods is given.

3.2.1. AESNN1 protein mapping method

In this protein mapping method, protein structure information is trained with a machine learning algorithm to create the numerical representations for protein sequences [28]. The numerical values of this method are given in Table 1.

3.2.2. Atchley factors protein mapping method

Atchley factors were proposed to eliminate metric problems encountered in protein sequences [29]. There are five factors in total in this method: bipolar (bp), secondary structure (ss), molecular volume (mv), relative amino acid composition (raac) and electrostatic charge (ec). Since there are 5 factors, it is a 5-dimensional method. Protein sequences are mapped using these five factors. Table 2 provides Atchley factors of amino acids.

3.2.3. Hydrophobicity protein mapping method

In the hydrophobicity method, the mapping processes of proteins are carried out according to the hydrophilic and hydrophobic information of the protein sequences in the polypeptide chains [30]. The hydrophobicity properties of amino acids are frequently used as an efficient way to compare and analyze amino acid sequences. Hydrophobicity protein mapping method consists of 1-dimensional numerical data. Hydrophobicity values of amino acids are given in Table 3.

3.2.4. Meiler parameters protein mapping method

Meiler parameters have been proposed to express reduced representations of amino acids [31]. Just like the Atchley factors, this method has many parameters. These parameters are steric (s), polarizability (p), volume (v), hydrophobicity (h), isoelectric point (ip), helix probability (hp), and sheet probability (sp). With this method, a 7-dimensional data is obtained when the protein sequences are mapped. Meiler parameters of amino acid codes are given in Table 4.

3.2.5. EIIP protein mapping method

The EIIP method has been proposed to predict interactions between proteins and between protein and DNA [32]. With this method, firstly, sequences are converted into genomic signals by Fourier transform. The power spectrum values are then calculated, and the amino acids are mapped. EIIP values of amino acids are given in Table 5.

3.2.6. CPNR protein mapping method

The CPNR protein mapping method was proposed to compare protein functions [33]. In this method, protein sequences were mapped according to codons. In the first step of the mapping process, amino

| Table 1 | AESNN1 values of amino acid codes. |
|---------|----------------------------------|
| Amino Acid Code | AESNN1 Value | Amino Acid Code | AESNN1 Value |
| A | 0.99 | L | 0.99 |
| R | 0.28 | K | 0.63 |
| N | 0.77 | M | 0.80 |
| D | 0.74 | F | 0.87 |
| C | 0.34 | P | 0.99 |
| Q | 0.12 | S | 0.99 |
| E | 0.59 | T | 0.42 |
| G | 0.79 | W | 0.13 |
| H | 0.98 | Y | 0.59 |
| I | 0.77 | V | 0.99 |

A protein sequence P (S) = \{A R N D C Q \ldots\} is mapped as C (S) = \{-0.99 0.28 0.77 0.74 0.34 0.12 \ldots\} by the AESNN1 protein mapping method.
Table 2

| Amino Acid Code | Hydrophobicity Value | Amino Acid Code | Hydrophobicity Value |
|----------------|----------------------|----------------|----------------------|
| A              | 1.8                  | M              | 1.9                  |
| C              | 2.5                  | N              | -3.5                 |
| D              | -3.5                 | P              | -1.6                 |
| E              | -3.5                 | Q              | -3.5                 |
| F              | 2.8                  | R              | -4.5                 |
| G              | -0.4                 | S              | -0.8                 |
| H              | -3.2                 | T              | -0.7                 |
| I              | 4.5                  | V              | 4.2                  |
| K              | -3.9                 | W              | -0.9                 |
| L              | 3.8                  | Y              | -1.3                 |

A protein sequence P(S) = [A R ...] is mapped as C(S) = \([-0.591 -1.302 -0.733 1.570 -0.146]\) by the hydrophobicity protein mapping method.

Table 3

| Amino Acid Code | CPNR Value | Amino Acid Code | CPNR Value |
|----------------|------------|----------------|------------|
| A              | 6.13       | M              | 0.0946     |
| N              | 0.828      | W              | 0.0548     |
| D              | 1.050      | F              | 0.0946     |
| C              | -1.343     | Y              | 0.0516     |
| E              | 1.239      | P              | 0.0198     |
| Q              | 0.931      | C              | 0.0829     |
| G              | -0.384     | T              | 0.0941     |
| H              | 0.336      | V              | 0.0057     |
| L              | -1.019     | L              | 0.0057     |

A protein sequence P(S) = [A R ...] is mapped as C(S) = \([-0.591 -1.302 -0.733\) 1.570 -0.146]\) by the Meiler parameters protein mapping method.

Table 4

| Amino Acid Code | C P N R Value | Amino Acid Code | C P N R Value |
|----------------|---------------|----------------|---------------|
| A              | 1.28          | 0.05          | 1.00          | 0.71          | 0.31          | 6.11          | 0.42          | 0.23          |
| R              | 2.34          | 0.29          | 6.13          | -1.01         | 10.74         | 0.36          | 0.25          |                |
| N              | 1.60          | 0.03          | 1.37          | 0.32          | 5.94          | 0.32          | 0.42          |                |
| D              | 1.46          | 0.22          | 4.63          | 1.23          | 5.71          | 0.38          | 0.32          |                |
| C              | 1.23          | 0.22          | 4.63          | 1.23          | 5.71          | 0.38          | 0.32          |                |
| E              | 1.23          | 0.22          | 4.63          | 1.23          | 5.71          | 0.38          | 0.32          |                |
| Q              | 2.94          | 0.29          | 5.89          | 1.79          | 5.67          | 0.30          | 0.38          |                |
| G              | 2.67          | 0.00          | 2.72          | 0.72          | 6.80          | 0.13          | 0.34          |                |
| H              | 1.89          | 0.22          | 4.77          | -0.99         | 9.99          | 0.32          | 0.27          |                |
| I              | 2.55          | 0.19          | 4.00          | 1.70          | 6.04          | 0.39          | 0.31          |                |
| K              | 1.89          | 0.22          | 4.77          | -0.99         | 9.99          | 0.32          | 0.27          |                |
| L              | 2.55          | 0.19          | 4.00          | 1.70          | 6.04          | 0.39          | 0.31          |                |

A protein sequence P(S) = [A R ...] is mapped as C(S) = \([-1.28 0.05 1.00 0.31 6.11 0.42 0.23]\) by the Meiler parameters protein mapping method.

Table 5

| Amino Acid Code | EIIP Value | Amino Acid Code | EIIP Value |
|----------------|------------|----------------|------------|
| A              | 0.0823     | M              | 0.0761     |
| R              | 0.0548     | W              | 0.0829     |
| N              | 0.0946     | F              | 0.0373     |
| D              | 0.0516     | Y              | 0.0036     |
| C              | 0.0829     | P              | 0.0559     |
| Q              | 0.0941     | T              | 0.0001     |
| G              | 0.0242     | H              | 0.1263     |
| H              | 0.0057     | V              | 0.0058     |
| L              | 0.0057     | L              | 0.0057     |

A protein sequence P(S) = [M W F Y P C ...] is mapped as C(S) = \(0.0823 0.0548 0.0946 0.0516 0.0198 0.0829\) by the EIIP protein mapping method.

3.2.7. Binary encoding protein mapping method

In the binary encoding mapping method, amino acids in protein sequences are represented by the values "0" and "1". The most widely used binary coding method is the one-hot coding method. In this method, each of the twenty standard amino acids is mapped with twenty-dimensional binary vectors. In the mapping process, amino acids are listed alphabetically, and numerical values are assigned to these amino acid codes. Binary encoding values of amino acid codes are given in Table 6.

3.2.8. PAM250 protein mapping method

The PAM250 protein mapping method is used to identify similarities in protein sequences by scoring aligned peptide sequences [34]. Values in the PAM250 matrix were obtained from comparison of aligned protein sequences with known homology and identification of observed point mutations. PAM250 values of amino acids are given in Table 7.

3.2.9. BLOSUM62 protein mapping method

BLOSUM matrices are used to score alignments between evolutionarily dissimilar protein sequences. It is a location-independent method and was first proposed in 1992 [35]. There are different types of matrices such as BLOSUM80, BLOSUM45. The numbers next to the matrix represent the similarity value. The similarities between the two organisms closest to each other in terms of evolution receive high score values. Therefore, the higher the score in BLOSUM matrices, the greater the similarity. For example, the BLOSUM80 matrix is used when aligning the protein sequences of two species closely related organisms. Conversely, the proteins of two organisms that are far apart in species can be aligned with the BLOSUM45 matrix. However, when the distance or proximity of organisms is not known exactly, the BLOSUM62 matrix is used. BLOSUM62 values of amino acids are given in Table 8.

Table 6

| Amino Acid Code | CPNR Value | Amino Acid Code | CPNR Value |
|----------------|------------|----------------|------------|
| M              | 1          | Q              | 29         |
| W              | 2          | S              | 31         |
| F              | 3          | A              | 37         |
| Y              | 5          | N              | 41         |
| P              | 7          | G              | 43         |
| C              | 11         | R              | 47         |
| H              | 13         | T              | 53         |
| T              | 17         | D              | 59         |
| V              | 19         | E              | 61         |
| L              | 23         | K              | 67         |

A protein sequence P(S) = [M W F Y P C ...] is mapped as C(S) = \([1 2 3 5 7 11 \ldots]\) by the CPNR protein mapping method.
A protein sequence $P(S) = [A C D \ldots]$ is mapped as $C(S) = [10000000000000000000 010000000000000000 00100000000000000000]$ by the binary encoding protein mapping method.

A protein sequence $P(S) = [A R N D C Q \ldots]$ is mapped as $C(S) = [10000000000000000000 010000000000000000 00100000000000000000]$ by the BLOSUM62 protein mapping method.

A protein sequence $P(S) = [A R N D C Q \ldots]$ is mapped as $C(S) = [2 6 2 4 4 4 \ldots]$ by the FIBHASH protein mapping method.

A protein sequence $P(S) = [A R N D C Q \ldots]$ is mapped as $C(S) = [0.001461 0.009875 0.000962 0.000531 0.000254 0.000646 0.000843 0.000099 0.000131 0.000680 0.0001445]$ by the Miyazawa potentials protein mapping method.

A protein sequence $P(S) = [A R N D C Q \ldots]$ is mapped as $C(S) = [M A K A T G R N N L V S P K K \ldots]$ by the FIBHASH protein mapping method.

A protein sequence $P(S) = [A C D E F G \ldots]$ is mapped as $C(S) = [0.07 0.09 0.12 0.09 0.0318 0.05109 0.031655 0.031328 0.00269 0.000782 0.005109 0.0031813]$ by the Micheletti potentials protein mapping method.
6

3.2.14. AVL-tree protein mapping method

In the AVL-tree method, the mapping of proteins is performed based on the AVL tree [40]. In this method, amino acid codes are first sorted alphabetically and placed in the AVL tree, respectively. After the necessary addition and balancing processes are completed, the depth values of the amino acid codes in the AVL tree are obtained and mapping is performed. Amino acid values of the protein mapping method developed with AVL-tree are given in Table 13.

3.2.15. Z-scale protein mapping method

The Z-scale protein mapping methods include the physicochemical properties of amino acids. It consists of NMR (Nuclear Magnetic Resonance) data and TLC (Thin Layer Chromatography) data. There are 5 parameters in total in this method. These are lipophilicity, polarity, electronegativity and electrophilicity. In this method, the numerical expressions of the protein sequences vary. A protein sequence P (S) = [M A K A T G R N L V S P K K ...] is mapped as C(S) = [0.72 -0.50 -0.53 0.09 0.55] by the Z-scale protein mapping method.

3.3. Designing a deep learning model

Deep learning is a type of machine learning and is used effectively today. One of the biggest reasons for its popularity is that data can be obtained more quickly and easily and the hardware needs that can analyze this data can be met [41]. Furthermore, being successful in complex and large data sets is also a factor in this sense. The biggest advantage of deep learning over machine learning is the process of feature extraction. While key features are manually extracted in machine learning-based approaches, there is an adaptive approach in deep learning [42]. The large number of data or the complexity of the data set make it difficult and even time consuming to extract features manually [16]. Such advantages have enabled deep learning to be used in almost every field. With deep learning, there are studies in the fields of biomedicine [43,44], bioinformatics [45], object recognition [46], robotics [47], energy [48]. In this study, recurrent neural networks, one of the deep learning models, were used and BiLSTM (Bidirectional Long/Short Term Memory) was designed. LSTM (Long/Short Term Memory) is more effective in sequential structures such as time and text series [49]. In the structure of LSTM, information from the past is used. Memory (Long/Short Term) is more effective in sequential structures such as time and text series. Therefore, the use of information in two directions provides an advantage in processing sequential data and time series. It consists of two LSTM units. One of them takes the input and processes it forward, while the other processes the information backwards. The parameters of the BiLSTM network designed in this study can be summarized as follows:

- Protein sequences mapped by each protein mapping technique were used in the input layer.
- BiLSTM was used in the second layer and 64 units were evaluated for this purpose. ReLU (Rectified Linear Unit) was used as the activation function.
- In the third layer, BiLSTM was used again, and 32 units were evaluated in this layer. ReLU is used as activation function again.
- Then, the outputs were flattened, and the data was made 1-dimensional.
- Then, to ensure that all data are in the same range, normalization was performed, and batch normalization was used.
- Finally, three fully connected layers are used and 240, 120 and 60 neurons are considered, respectively.
- In the last layer, classification was made, and 2 neurons were used in this layer to determine whether there was an interaction. Softmax calculation was used as the activation function.
- Categorical cross-entropy was used to determine the loss of the model.
- SGD (Stochastic Gradient Descent) was used for the optimization process.
- The training was carried out with 500 epochs.
- The performances of the protein mapping methods were determined by 10 fold cross-validation.
- All parameters used in the study were determined by trial-and-error approach and the parameters that gave the best results were evaluated.

The flow chart of the study is given in Fig. 2.

4. Application results

In the study, the performances of protein mapping methods were evaluated and compared after the classification process. Accuracy, F1-score, specificity, sensitivity, and AUC values were calculated to compare the performances of protein mapping methods. Average classification results are given in Table 14.

When the results given in Table 2 were examined, it was observed that the most effective result was obtained with the entropy-based method. With this method, an accuracy of 88.94% was obtained. Then the most effective method was the Z-scale method. The accuracy score of this method was calculated as 91.23%. Finally, the AVL-tree method showed a success rate of over 90% and the accuracy score was 90.44%. Other than these methods, no protein mapping method has shown a success rate of over 90%. While 89.33% accuracy was obtained with the FIBHASH protein mapping method, 88.94% accuracy was calculated with the AESNN1 method. The CPNR protein mapping method showed the closest performance to these methods, and an average of 88.52% accuracy was obtained with this method. In addition to these methods, accuracy values of 87.63%, 86.20%, 85.54%, 82.58% and 81.10% were obtained with the methods of PAM250, binary encoding, BLOSUM62, Atchley factors and Meiler parameters, respectively. The remaining protein mapping methods showed accuracy below 80%. EIP and hydrophobicity methods produced close scores, and accuracy scores of 78.89% and 79.65% were observed, respectively. While an accuracy score of 75.39% was obtained with Micheletti potentials, this rate increased in Miyazawa energies and rose to 77.31%. AUC score is a very important evaluation criterion in bioinformatics and medical studies [50-52]. A classifier is considered a reasonable classifier if its AUC score is greater than 0.7 [53]. In addition, values between 0.8 and 0.9 are expressed as a great classifier, while a value greater than 0.9 indicates that the classifier has excellent performance [54]. When a comparison is

| Table 13 |
|----------|
| AVL-tree values of amino acid codes. |
| Amino Acid Code | AVL-tree Value | Amino Acid Code | AVL-tree Value |
| A | 4 | L | 3 |
| R | 3 | K | 2 |
| N | 0 | M | 4 |
| D | 4 | F | 4 |
| C | 3 | P | 3 |
| Q | 2 | S | 1 |
| E | 2 | T | 3 |
| G | 3 | W | 2 |
| H | 1 | V | 3 |
| I | 3 | V | 4 |

A protein sequence P (S) = [A R N D C Q ...] is mapped as C(S) = [4 3 0 4 3 2 ...] by the AVL-tree protein mapping method.
made according to the AUC scores, only the AVL-tree, entropy-based and Z-scale methods showed a value of over 90%. Therefore, these methods have performed an excellent classification. Apart from these methods, almost every method has performed a great classification process. However, since the AUC scores of hydrophobicity and Micheletti potentials were below 0.80, it was acceptable. Fig. 3 shows ROC graphs of protein quantification methods with AUC scores above 0.9.

5. Discussion

In this study, viral-host interactions between SARS-CoV-2 virus and human proteins were analyzed by various protein mapping methods. Therefore, the performances of these methods were compared, and it was explained why the results differed according to the protein mapping methods. Micheletti potentials and Miyazawa energies are structure-based methods. The reason why these two methods are the most ineffective may be that the structure information of protein sequences is not needed in the study. For this reason, key feature may have been lost.
Obtaining structure-based features will increase the performance of these methods. Looking at the results in Table 2, it is seen that physicochemical-based methods are also generally ineffective. Although Atchley factors, hydrophobicity, Meiller parameters and EIIP in this category performed better than structure-based ones, they were generally ineffective. The main reason for this situation is that the mapping process requires a certain information, just like structure-based methods. In physicochemical-based methods, chemical information of proteins is needed. Since no chemical information was analyzed in this study, information loss may occur. When the performances of evolution-based methods (PAM250 and BLOSUM62) are examined, it is seen that protein mapping methods in this category are more successful than structure-based and physicochemical-based methods. Evolution-based methods are often used in alignment studies [55,56]. Performing the classification process after the alignment process in the study can significantly improve the performance of these methods. Character-based methods seem to be more successful. The CPNR, binary encoding, and FIBHASH methods all achieved an accuracy score of over 85%. The reason behind this is that there is no need for specific information while mapping with these methods. Therefore, there may be no loss of information. The second most effective method was the Z-scale method. Obtaining SAR (Structure Activity Relationship) features with this method has been instrumental in the success of the method. The Z-scale method is also a physicochemical-based method [57]. However, it was more effective than other physicochemical-based methods (Atchley factors, hydrophobicity, Meiller parameters and EIIP). In addition to these, it is seen that the most effective method is the entropy-based method. The reason why these two methods are successful is that they have a dynamic structure. In all other methods, protein sequences have a static structure. In a static structure, amino acids in protein sequences always take the same numerical values. However, since there is an adaptive approach in the entropy-based and Z-scale method, protein sequences can take different values. In the entropy-based method, the numerical values of the amino acid codes vary according to the length of the protein sequence and the repetition frequency of the amino acids in the sequence. There are 5 parameters in the Z-scale method, and this calculation process varies according to the protein sequence and the order of the amino acid codes in the sequence. For these reasons, these two methods are adaptive methods. Additionally, in some protein mapping methods, different amino acids take on the same value. For example, in the AESNN1 protein mapping method, A and P amino acids take the same values. Similarly, amino acids L and I take the same values in the EIIP protein mapping method. In this and similar methods, degeneration is observed when different amino acids take the same numerical expressions [58]. This causes the methods to be ineffective.

6. Conclusion

In this study, viral-host interactions between SARS-CoV-2 and human proteins were predicted by computational-based approaches. The study consisted of four different stages. In the first stage, viral-host interactions between SARS-CoV-2 virus and human proteins were obtained and a positive data set was created. In addition, non-interacting protein sequences were obtained, and a negative data set was created. In the second stage, the protein sequences in both negative and positive data sets were mapped with various protein mapping methods and made ready for classification by deep learning algorithm. In the third stage, a deep learning model was designed and the BiLSTM algorithm was used. At the last stage, the performances of protein mapping methods were determined by accuracy, F1-score and AUC scores and their performances were compared. The most ineffective protein mapping method was Michielleti potentials, with an accuracy of 75.39%. The most effective classification process was performed with the entropy-based protein mapping method, and an accuracy of 94.74% was obtained with this method. When the results were examined in general, it was determined that the most ineffective methods were structure-based methods. Furthermore, it has been observed that the most effective method for physicochemical-based methods is Atchley factors. With this method, an accuracy of 82.58% was obtained, while an AUC score of 0.83 was calculated. Subsequently, the most effective methods have been evolution-based methods. Among these methods, the PAM250 matrix achieved an accuracy score of 87.63% by performing a successful classification process, while it achieved an AUC score of 0.87. The fourth most effective method was obtained by AESNN1 protein mapping method, which is one of the machine learning-based methods. With this method, an accuracy score of 88.94% was calculated, while an AUC score of 0.81 was achieved. The third most effective mapping method was obtained with FIBHASH, one of the character-based protein mapping methods. With this method, an accuracy score of 89.33% was obtained, while an AUC of 0.88 was calculated. On the other hand, AVL, which is an algorithm-based method, performed the second most efficient classification process. The accuracy score of this method was calculated as 90.44%. The AUC score was determined as 0.91. However, there are some limitations of the developed study. According to the application results, it has been determined that computational-based approaches can be used effectively in this field. There are not many studies in this topic in the literature. Therefore, the study needs to be supported by various other approaches (other mapping methods, alignment algorithms, etc.). Furthermore, finding new viral-host interactions between SARS-CoV-2 and human proteins will increase the number of data. Therefore, the results obtained in this study can be further improved. The large number of data will enable the deep learning model to learn better and the results will be more reliable. In addition to these, with different deep learning and machine learning algorithms, the results can be evaluated in a different way. To preserve the diversity and performance of the study, it should also be evaluated with other deep learning algorithms. Apart from the protein mapping methods used in this study, the use of different protein mapping approaches is important for the accuracy and reliability of the study. It is important to evaluate the study with other protein mapping methods in this regard. When the results were examined in general, it was observed that all protein mapping methods performed a successful classification process with the designed deep learning model. In this way, the use of computational-based approaches in this field will help both to better understand the structure of the virus and to determine the result of its interactions with human proteins more quickly.

Authorship contributions

Conception and Design of the Study: T. B. Alakus.
Acquisition of Data: T. B. Alakus.
Analysis and/or Interpretation of Data: T. B. Alakus, I. Turkoglu.
Drafting the Manuscript: T. B. Alakus.
Revising the Manuscript: T. B. Alakus, I. Turkoglu.
Approval of the Version of the Manuscript to be Published: T. B. Alakus, I. Turkoglu.

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.

Acknowledgements

I want to thank my wife who has been with me in every moment of my life and I want to make her a proposal. Dilan will I marry you?
[55] S.F. Altschul, Amino acid substitution matrices from an information theoretic perspective, J. Mol. Biol. 219 (3) (1991), 555-565, https://doi.org/10.1016/0022-2836(91)90193-X.

[56] G.M. Streiemer, A. Akoglu, Sequence alignment with GPU: performance and design challenges, in: Proceedings in International Symposium on Parallel and Distributed Processing, 2009, https://doi.org/10.1109/IPDPS.2009.5161066, 23-29 May, Italy, Rome.

[57] M. Sandberg, L. Eriksson, J. Jonsson, M. Sjostrom, S. Wold, New chemical descriptors relevant for the design of biologically active peptides. A multivariate characterization of 87 amino acids, J. Med. Chem. 41 (14) (1998) 2481-2491, https://doi.org/10.1021/jm970057s.

[58] S.S.T. Yau, J. Wang, A. Niknejad, C. Lu, N. Jin, Y.K. Ho, DNA sequence representation without degeneracy, Nucleic Acids Res. 31 (12) (2003) 3078-3080, https://doi.org/10.1093/nar/gkg632.