Open Peer Review on Qeios

RESEARCH ARTICLE

Evaluation of Molecular Docking by Deep Learning and Random Forests: A Hybrid Approach Based on Pseudo-Convolutions

Janderson Romário Borges da Cruz Ferreira[1], Allan Rivalles Souza Feitosa[2], Juliana Carneiro Gomes[3], Abel Guilhermino da Silva-Filho[2], Wellington P. dos Santos[3,1]

1 Escola Politécnica de Pernambuco, Universidade de Pernambuco, Brazil

2 Centro de Informática, Universidade Federal de Pernambuco, Brazil

3 Departamento de Engenharia Biomédica, Universidade Federal de Pernambuco, Brazil

Funding: The authors thank the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, and Instituto do Complexo Econômico-Industrial da Saúde, iCEIS, Brazil, for the partial financial support for this research.

Potential competing interests: No potential competing interests to declare.

Abstract

Purpose: Molecular docking prediction plays a pivotal role in intelligent drug design, offering significant advantages in the development of antivirus medications and vaccines. By accurately evaluating the interactions between drug molecules and target proteins, researchers can effectively expedite the discovery and development of vital pharmaceutical solutions, aiding in the mitigation of epidemics, pandemics, and the urgent need for improved vaccines.

Methods: We propose an intelligent hybrid architecture for estimating molecular docking between two proteins using deep networks based on pseudo-convolutions and Random Forests. As input, we used the sequences of characters representing the bases of each protein. Through two layers of pseudo-convolutions, these sequences are reorganized and, finally, represented as co-occurrence matrices, concatenated into a feature vector. To validate our proposal, we used a combination of the public datasets Affinity Benchmark 3 and Negatome 2[1][2].

Results: The experimental results revealed that the hybrid architecture, comprising two layers of pseudo-convolution followed by a decision layer with an ensemble of 25 Random Forests, yielded average accuracy of 94%, AUC of 93%, sensitivity of 94%, and specificity of 78%. This demonstrates the feasibility of achieving robust estimates of molecular docking without relying on 3D molecule modeling.

Conclusion: These findings the potential of rapidly estimating protein affinity, providing valuable insights into drug interactions, molecular dynamics, and facilitating the intelligent design of pharmaceuticals, encompassing vaccines and antiviral drugs. These advancements play a critical role in cost reduction related to laboratory analysis and expediting the timely delivery of solutions to both the market and society at large, particularly in the context of epidemic outbreaks, pandemics, and the urgent demand for supplementary vaccines.

Allan Rivalles Souza Feitosa, Juliana Carneiro Gomes, Abel Guilhermino da Silva-Filho, and Wellington Pinheiro dos

Santos equally contributed to this work.

Corresponding author: Wellington Pinheiro dos Santos, wellington.santos@ufpe.br

1. Introduction

1.1. Motivation and problem characterization

The Fourth Industrial Revolution, characterized by the integration of advanced technologies, has the potential to significantly impact the development of nations in the Global South, particularly in the fields of Biomedical Engineering, digital health, and the pharmaceutical industry[3][4][5][6][7][8][9][10][11]. These technologies, such as Artificial Intelligence (AI), the Internet of Things (IoT), and big data analytics, offer unique opportunities for improving healthcare access, delivery, and outcomes in resource-constrained settings[12][13][14][15][16][17]. In Biomedical Engineering, these advancements can enhance the design and development of medical devices, diagnostics, and prosthetics, providing affordable and tailored solutions to address specific healthcare needs[18][19][20]. Digital health platforms can revolutionize healthcare delivery by enabling remote monitoring, telemedicine, and personalized health management, bridging the gap between patients and healthcare providers, especially in remote or underserved areas[21][22][23]. Furthermore, the pharmaceutical industry can benefit from intelligent drug design and discovery through computational methods that leverage AI and machine learning. These techniques enable faster and more efficient identification of potential drug candidates, reducing costs and time associated with traditional methods[24][25][26][27][28][29][30][31]. The integration of these technologies holds immense potential for the Global South, offering opportunities to leapfrog and overcome existing healthcare challenges, leading to improved healthcare access, better patient outcomes, and enhanced pharmaceutical development.

Estimating the affinity between two proteins holds significant importance for the health and pharmaceutical industries. Protein-protein interactions play a central role in numerous biological processes and disease pathways, making them attractive targets for therapeutic intervention. Accurate estimation of protein affinity provides insights into the strength and specificity of these interactions, enabling the identification of potential drug targets and the design of effective therapeutic strategies[32][33][34][35][36][37]. Precise knowledge of protein affinity facilitates rational drug design, lead optimization, and the development of personalized medicine approaches[32][33][34][35][36][37]. Moreover, it aids in understanding the mechanisms underlying various diseases, paving the way for the discovery of novel treatment options[37].

The reduction of time and cost associated with molecular docking poses significant challenges in the field of drug discovery[34][38][39][40][41][42][43][44][45][46][47][48][49][50]. Molecular docking involves simulating the binding between a small molecule and a target protein, a process that necessitates extensive computational resources and experimental validation. The exploration of the vast conformational space and the accurate prediction of protein-ligand interactions contribute to the time-consuming nature of this approach. These limitations hinder the efficiency of the drug discovery process and increase costs through the need for extensive experimental iterations[34][38][39][40][41][45][49][50]. Overcoming these difficulties requires innovative approaches that can expedite molecular docking while maintaining accuracy and reducing

the reliance on resource-intensive experimental procedures.

Computational intelligence, more specifically machine learning, offers valuable advantages in estimating the degree of molecular docking, addressing the challenges mentioned above. By harnessing large-scale datasets, these techniques enable the development of predictive models that can learn complex patterns and relationships between protein structures and binding affinities[51][52][53][54][55][56][57]. Machine learning models can capture non-linear interactions and subtle features that are challenging to model using traditional methods. Through the integration of various input features such as protein sequences, structural information, and physicochemical properties, these models can accurately predict protein affinity[58][59][60][61][62][63][64]. The application of machine learning and computational intelligence streamlines the screening of large chemical libraries, accelerating the identification of potential drug candidates[58][59][60][61][62][63][64].

In recent years, there have been significant advancements in computational models based on machine learning for estimating molecular docking[65][66][67][68]. The research field has witnessed the development and refinement of various approaches utilizing deep learning architectures such as convolutional neural networks (CNNs), recurrent neural networks (RNNs), and graph neural networks (GNNs). These models have demonstrated impressive performance in support diagnosis applications based on biomedical images, signals and clinical parameters, like in breast cancer, Covid-19, mental disorders, and in neurodegenerative diseases like Alzheimer's and Parkinson's[69][70][71][72][73][74][75][76][77][78][79][80][81][82][83][84][85][86][87][88]. Such models have been used successfully in accurately predicting protein-protein binding affinities. Furthermore, the availability of extensive protein structure and interaction databases has facilitated the creation of comprehensive and robust machine learning models. Integrating multiple data modalities, including protein sequences, structures, dynamics, and ligand information, has further enhanced the predictive power of these models[65][66][67][68][89][90][91]. In addition, transfer learning approaches have emerged, where pre-trained models on related tasks are fine-tuned for molecular docking prediction, leading to efficient knowledge transfer and faster model convergence. The state-of-the-art computational models based on machine learning hold significant promise in accurately estimating molecular docking, ultimately contributing to the advancement of drug discovery and the development of effective therapeutics[65][66][67][68][89][90][91].

Building intelligent algorithms for molecular docking that avoid the 3D modeling of proteins by angles and energies holds great importance in the field of computational drug discovery. By relying solely on protein descriptions based on strings of characters representing sequences of amino acids, these algorithms offer significant advantages and possibilities. One of the key benefits is the reduction in computational complexity and time required for 3D modeling, which can be resource-intensive and time-consuming. By directly utilizing protein sequences, intelligent algorithms can leverage the wealth of information encoded within them, including structural motifs, functional domains, and evolutionary relationships. This approach not only accelerates the molecular docking process but also opens up new avenues for exploring a wide range of protein-protein interactions. Additionally, focusing on protein sequence-based descriptions allows for broader applicability, as it enables the analysis of proteins with unknown or uncharacterized structures. Moreover, these algorithms can leverage machine learning techniques to learn and predict the protein-protein interactions solely from sequence information, further enhancing their accuracy and potential. Therefore, the development of intelligent algorithms that utilize protein descriptions based on amino acid sequences represents a promising direction in molecular docking research,

enabling efficient and accurate estimation of protein affinity while bypassing the challenges associated with 3D modeling.

In this work, we propose an intelligent hybrid architecture for estimating molecular docking between two proteins using deep networks based on pseudo-convolutions and Random Forests. Pseudo-convolution networks consist of an iterative process where, given a given number of layers, an input sequence representative of a pair of candidate-target proteins, represented as a sequence of characters, is broken into smaller parts with size in power of two. In the final layer, the set of sequences obtained is represented by co-occurrence matrices that model neighborhood relationships and character populations. The matrices corresponding to each character segment of the output are transformed into vectors and these are then concatenated, forming a feature vector. The vectors thus represented are classified by a Random Forest. The molecular docking estimation problem was modeled as a classification problem. This approach was inspired by[92]: they improved the virus identification results obtained from RT-PCR devices by training machine learning algorithms over training a large dataset: 347,363 virus DNA sequences from 24 virus families and SARS-CoV-2. The authors obtained results for sensitivity and specificity from 97% to 99%, demonstrating that the molecular diagnosis of Covid-19 can be optimized by combining RT-PCR and the pseudo-convolutional method to identify DNA sequences for SARS-CoV-2 with larger specificity and sensitivity values.

1.2. Related Works

The process of assessing the potential bonding between a ligand and a target protein is a multifaceted task that involves the selection of candidates from extensive databases for subsequent *in vitro* and *in vivo* validation[93]. This process finds applications in various fields such as drug discovery, peptide generation, DNA ligand discovery, and more.

Virtual screening is a technique employed to select ligand candidates for target proteins, with the aim of reducing the number of compounds that need to undergo *in vitro/vivo* testing phases[93]. This approach helps to minimize the associated costs of these processes. Over the years, machine learning (ML)-based algorithms have been proposed for virtual screening. Scientific literature showcases a range of methods, including classical ML techniques like K-nearest neighbors and gradient boosting[94], as well as advanced graph learning methods[95], which are trained using large ligand-protein datasets tailored for supervised learning.

[96] exemplify the process of discovering new umami peptides for the food industry using machine learning and molecular docking. They successfully identified six novel peptides derived from lamb bones and determined that hydrogen bonding and electrostatic interactions were the primary forces involved. Their work employed a deep learning approach based on a neural network model that combined fully connected neural networks (MLP) and recurrent neural networks (RNN) to predict umami taste and evaluate the umami taste threshold of unknown peptides.

Addressing the limited exploration of docking methodologies for the interactions between nucleic acids and DNA intercalating agents,[97] proposed a machine learning method to predict changes in DNA melting temperature upon drug binding. They compared their approach with Autodock, Dock6, and Consensus methods.

In a drug repurposing study involving FDA-approved drugs,[98] identified Cobimetinib as an A-FABP inhibitor. They

investigated a dataset of approximately 2,600 compounds and employed a ligand-based machine learning approach and a structure-based molecular docking method, both based on Naive Bayesian. Additionally, they utilized t-SNE for data visualization.

[99] developed an approach to propose novel drug-like molecules by combining a global molecular property optimization strategy with a machine learning algorithm for docking score prediction. Their method generated numerous novel molecules with high docking scores for a specific protein. The machine learning model served as part of the objective function to evaluate the docking score of the proposed solution candidates.

In general, machine learning techniques have been applied to target-specific scoring tasks. Various studies, such as those by[94][100][101], have demonstrated the application of different classical machine learning approaches, including Logistic Regression, Gradient Boosting, K-Nearest Neighbor, Support Vector Machine, Random Forest, and Multi-Layer Perceptron.

Most of the computational approaches for protein docking found in scientific literature require the simulation of the complex in 3D prior to predicting scores or determining active/inactive compounds[102][103][104][105][106][107][108][109][110][111]. In this study, we propose a solution based on pseudo-convolution for protein representation and protein docking prediction. This method eliminates the need for 3D complex simulation.

2. Materials an Methods

2.1. Datasets

In order to implement and validate our proposal, we utilized two datasets: Affinity Benchmark Version 3 (257)[1] and Negatome 2[2].

The Affinity Benchmark Version 3 is an updated version of the protein-protein docking benchmark 2. According to the authors[1], this update incorporates 40 new test cases, resulting in a 48% increase compared to Benchmark 2.0. The 124 unbound-unbound test cases in Benchmark 3.0 are categorized into three groups: 88 rigid-body cases, 19 medium-difficulty cases, and 17 difficult cases. The classification is based on the extent of conformational changes occurring at the interface upon complex formation. The expansion of Benchmark 3.0 not only provides the scientific community with a larger set of test cases for evaluating docking methods but also facilitates the development of new algorithms that require abundant training examples. For a visual demonstration of complexes from Affinity Benchmark Version 3, please refer to Table 1.

Complex	Cat.	PDB ID 1	Protein 1	PDB ID 2	Protein 2	I-RMSD (Å)	δASA(Å2)	BM version introduced
Rigid-body (162)								
1AHW_AB:C	AA	1FGN_LH	Fab 5g9	1TFH_A	Tissue factor	0,69	1899	2
1DQJ_AB:C	AA	1DQQ_CD	Fab Hyhel63	3LZT_	HEW lysozyme	0,75	1765	2
1E6J_HL:P	AA	1E6O_HL	Fab	1A43_	HIV-1 capsid protein p24	1,05	1245	2
1JPS_HL:T	AA	1JPT_HL	Fab D3H44	1TFH_B	Tissue factor	0,51	1852	2
1MLC_AB:E	AA	1MLB_AB	Fab44.1	3LZT_	HEW lysozyme	0,6	1392	2

Table 1. Example of the Affinity Benchmark 3 dataset organization

Negatome 2.0 introduced a significant methodological advancement through the implementation of an intricate text-mining procedure for manual annotation. A revised version of Excerbt, a sophisticated text-mining tool employing semantic sentence analysis, was utilized to identify potential non-interactions. Subsequent manual inspection revealed that nearly 50% of the text-mining results with the highest confidence values corresponded to pairs of non-interacting proteins (NIP). The expansion of the database in Negatome 2.0 is noteworthy, with an increase of over 300% compared to its previous iteration[2]. For a visual representation of complexes from Negatome 2, refer to Table 2.

Table 2. Example of Negatome 2 dataset organization

No.	Protein A	Protein B	PMID	Evidence
1	Q6ZNK6	Q9Y4K3	15047173	MI:0019 - coimmunoprecipitation
2	Q9NR31	Q15797	17356069	MI:0019 - coimmunoprecipitation
3	P11627	P53986	20155396	MI:0411 - enzyme linked immunosorbent assay
4	P33176	Q96EK5	16225668	MI:0059 - gst pull down
5	Q9NPY3	P02745	11994479	MI:0411 - enzyme linked immunosorbent assay

Since the objective of this study is to propose a method capable of predicting the presence or absence of docking between two proteins, the construction of an appropriate database necessitates the inclusion of complexes with docking (labeled as 1) as well as complexes without docking (labeled as 0). Unfortunately, existing public databases in the literature do not encompass both classes simultaneously. Consequently, a combination of selected databases was employed, each providing instances of the required classes. The utilized datasets comprise Affinity Benchmark 3, consisting of 257 complexes with docking, and Negatome 2, which encompasses 6184 complexes without docking. A subset of the final database is illustrated in Table 3.

Table 3. Example of the organization of the full dataset used in this work

Protein A	Protein B	Label
1FGN_LH	1TFH_A	1
1DQQ_CD	3LZT	1
1E6O_HL	1A43	1
1JPT_HL	1TFH_B	1
1MLB_AB	3LZT	1
Q6ZNK6	Q9Y4K3	0
Q9NR31	Q15797	0
P11627	P53986	0
P33176	Q96EK5	0
Q9NPY3	P02745	0

2.2. Proposed method

In this study, we introduce a novel method for extracting features from ribonucleic acid (RNA) sequences of proteins. Our approach, inspired by the pseudo-convolution machine utilized by[92], aims to represent protein RNAs as numerical feature vectors derived from co-occurrence matrices that capture the neighborhood characteristics within a molecule. This representation enables a machine-learning model to classify the presence or absence of docking between a pair of proteins. Notably, our method eliminates the necessity of simulating the 3D structure of the protein complex for docking classification. To validate our proposal, we employ the proposed dataset and utilize a random forest model. Further elaboration on the full database, random forest implementation, and evaluation metrics can be found in Subsections 2.1, 2.3, and 2.5, respectively. The general pipeline of our proposal is presented in Figure 1, and Figure 2 illustrates our specific contribution within the context of this research.

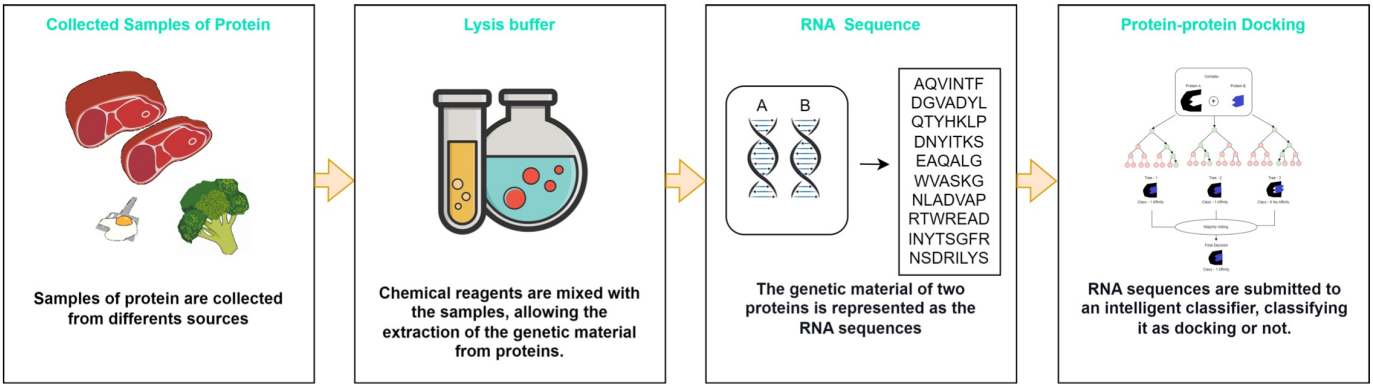

Figure 1. General scheme of the proposal: Initially, genetic material is acquired through sample collection, and subsequently, the RNA sequences are obtained and stored as text files. These RNA sequences undergo a pseudo-convolutional representation process, which involves their conversion into numeric vectors. Subsequently, a random forest model is employed to classify these representations into two categories: 0 or 1. Here, 1 signifies the presence of docking between the samples, while 0 indicates the absence of docking. The samples used in this study were sourced from two distinct datasets: the Affinity Benchmark 3 and the Negatome 2 dataset.

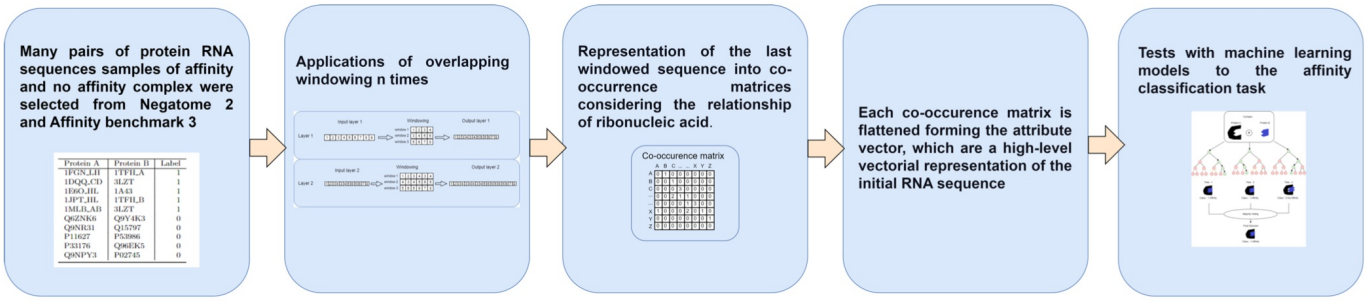

Figure 2. Detailed contribution: A total of 6,441 complexes, encompassing both docking and non-docking cases, were utilized for training the Random Forest models. To preprocess the RNA sequences, each sequence was concatenated and subjected to subsequent segmentation into sub-sequences, while considering windowing and overlapping parameters. These sub-sequences were further concatenated and had the potential to undergo additional segmentation based on the same windowing and overlapping parameters. Subsequently, the resulting concatenated sequence was represented as 26 × 26 co-occurrence matrices, which effectively captured the distribution of RNA neighborhoods. To create the feature vector for classification, the flattened form of the 26 × 26 matrix was employed as the input.

The process of feature extraction using a pseudo-convolutional machine can be delineated as follows. Initially, the RNA sequences are concatenated to form a complete sequence. This complete sequence is subsequently partitioned into n subsequences. To enable overlapping between these subsequences, the size of the overlapping segments is determined by a parameter passed to the method. Notably, all n sub-sequences are then concatenated, resulting in an extended sequence. This entire process can be repeated n times, generating increasingly extensive sequences with novel features at each iteration.

The final sequence is represented by a 26 by 26 co-occurrence matrix. This matrix captures the frequency of occurrences for specific pairs of ribonucleic acid (RNA) and provides insights into the proximity of RNA to their neighboring counterparts. For instance, if we consider the RNA sequence read from left to right, the current RNA element being analyzed is denoted as E, while its adjacent neighbor is labeled as P. Consequently, the corresponding matrix element located at the intersection of row E and column P is incremented accordingly to signify their association. For a detailed visual representation of this process, refer to Figure 3.

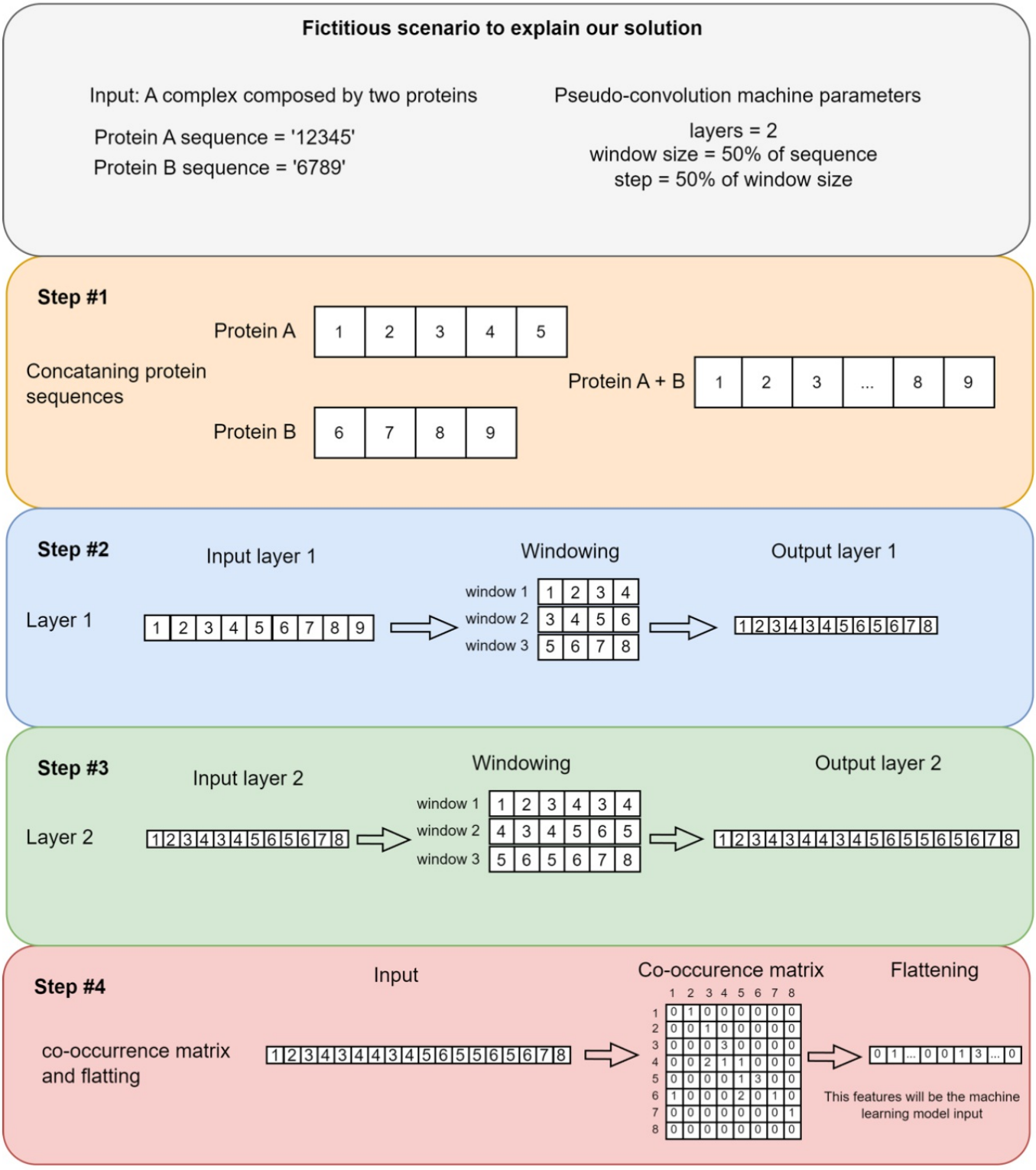

Figure 3. Steps of the proposed method: A new method of representing genome sequences has been developed, which involves analyzing the relationship between ribonucleic acids. First, RNA sequences are concatenated, then overlapping windowing is applied, resulting in small sequences. Successively, all small sequences are concatenated. The overlapping windowing can be reapplied n times using as input the concatenation of the small sequences. Next, the final concatenated sequence is represented by a 26 *por*26 co-occurrence matrix representing the distribution of nucleotide neighbors. Finally, the co-occurrence matrix is flattened to be used by a machine learning model.

Upon feature extraction from all complexes, the resulting database was partitioned into balanced train and test sets based

on the mean and standard deviation attributes. This rigorous splitting procedure is crucial to ensure robust model evaluation. The distribution of instances per class and dataset can be observed in Table 4.

Table 4. Number of instances of training and test sets

	No Docking	Docking	Total
Train set	4948	205	5153
Test set	1237	51	1288

2.3. Random Forest

The Random Forest algorithm is a widely adopted machine learning technique utilized for both classification and regression tasks. During the training process, this method constructs multiple decision trees. In classification scenarios, the output of the Random Forest model corresponds to the class that is most frequently selected by the constituent trees. Conversely, in regression scenarios, the output can be either the average prediction or the mean value generated by the individual trees[112][113][114].

Considering the nature of the problem investigated in this study, the Random Forest approach was specifically implemented to address the classification task at hand. This enabled the accurate categorization of protein complexes as either "docking" or "not docking". A visual representation of a Random Forest model designed for the classification of protein docking between two proteins is provided in Figure 4.

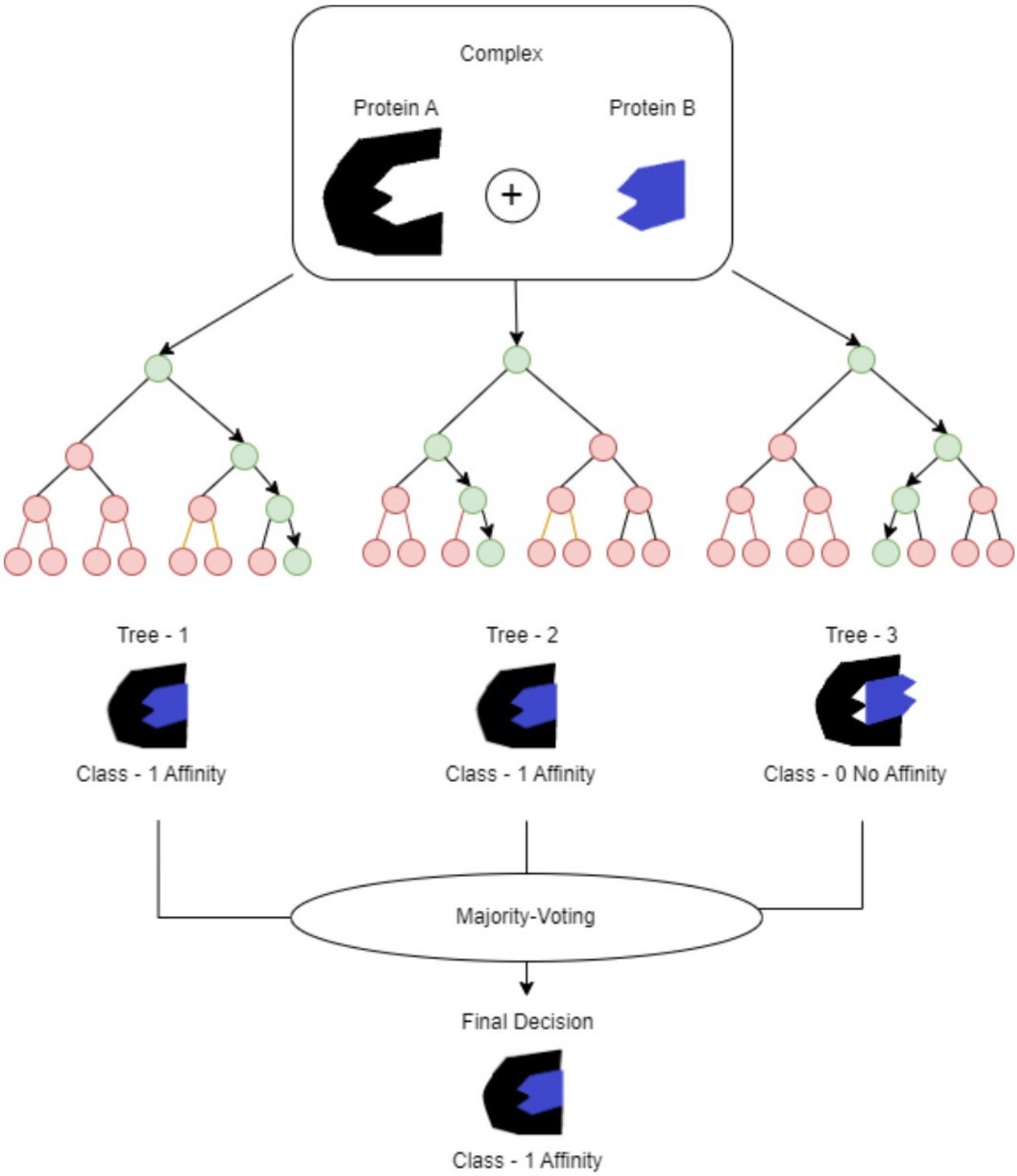

Figure 4. Random forest for classifying docking between two proteins. Trees 1 and 2 classified the complex as class 1(Docking), while the third tree classified it as class 0 (No docking). However due the majority voting the final decision was classify this complex as class 1.

2.4. Classification

In order to find a good random forest configuration to the docking classification task, ten different random forest configurations were evaluated, and the guide metric used to define the best was the kappa coefficient. Each configuration

was evaluated in the train set using cross-validation with k fold equal to 10. In addition, each configuration was executed 30 times, so the results found are statistically more reliable. Table 5 shows all random forests evaluated in this work. Finally, after finding the best random forest configuration, it was evaluated on the test dataset, and all proposed evaluation metrics were generated.

Table 5. All setups of random forests were used to find the best for the classification docking task. The number of trees varied between 100 to 1000.

ID	Number of Trees
RF_100	100
RF_200	200
RF_300	300
RF_400	400
RF_500	500
RF_600	600
RF_700	700
RF_800	800
RF_900	900
RF_1000	1000

To determine the optimal configuration for the Random Forest algorithm, a comprehensive evaluation was conducted involving cross-validation with 10 folds, repeated 30 times, resulting in a total of 300 experiments on the training set. Subsequently, in the testing phase, the best classifier identified during the validation step was trained on the entire training set and subsequently evaluated on the independent test set.

Notably, the training set exhibited a substantial class imbalance, with the number of protein sequence pairs categorized as "no docking" being 25 times greater than the number of pairs classified as "docking" (4948 instances vs. 205 instances, respectively, as can be seen on Table 4). This considerable imbalance can significantly impact the obtained results, despite the inherent robustness of Random Forest classifiers in handling class imbalance. Consequently, a second approach was devised to address this issue. The training set was divided into 25 balanced subsets, and each subset was employed to train an individual classifier with the previously determined optimal configuration. These 25 classifiers were then combined as an ensemble, and collectively assessed on the test set to enhance classification performance.

2.5. Evaluation Metrics

To evaluate objectively the classification results, we used the following methods: the κ index, the *overall accuracy*, the *confusion matrix*, the sensitity, the specificity, and the AUC. The *confusion matrix* for a universe of classes of interest $\Omega = \{C_1, C_2, \ldots, C_m\}$ is a $m \times m$ matrix $\mathbf{T} = [t_{i,j}]_{m \times m}$ where each element $t_{i,j}$ represents the number of objects belonging to class C_j but classified as C_i[76][115][116][117][118][119][120].

The *overall accuracy* is the probability that the experiment will provide correct results, that is, to correctly classify the pairs of protein sequences as having affinity or not. In other words, it is the probability of the true positives (TP) and true negatives (TN) among all the results. The sensitivity metric indicates the rate of true positive, while specificity is the rate of true negatives. AUC stands for Area under the ROC curve, and can be approximated by the mean. The ROC curve, in turn, is a graph showing the True positive rate vs. False positive rate. Finally, the Kappa index is a statistical correlation rate[115]. Thereby, Accuracy, Sensitivity, Specificity, AUC, and Kappa Index metrics can be calculated according to the equations [eq:accuracy], [eq:sensitivity], [eq:specificity], [eq:auc], and [eq:kappa] respectively.

$$\text{Accuracy} = \rho_v = \frac{n_{TP} + n_{TN}}{n_{TP} + n_{TN} + n_{FP} + n_{FN}},$$

$$\text{Sensitivity} = \rho_{TP} = \frac{n_{TP}}{n_{TP} + n_{FN}},$$

$$\text{Specificity} = \rho_{TN} = \frac{n_{TN}}{n_{TN} + n_{FP}},$$

$$\text{AUC} = \int_0^1 \rho_{TP} d\rho_{TN},$$

$$\kappa = \frac{\rho_v - \rho_z}{1 - \rho_z},$$

where

$$\rho_z = \frac{\sum_{i=1}^{m} \left(\sum_{j=1}^{m} t_{i,j} \right) \left(\sum_{j=1}^{m} t_{j,i} \right)}{\left(\sum_{i=1}^{m} \sum_{j=1}^{m} t_{i,j} \right)^2},$$

ρ_v is the accuracy, and $t_{i,j}$ is the element of the confusion matrix in position (i, j), i.e. the number of instances in the training set belonging to the i-th class but classified as belonging to the j-th class by the machine learning model under evaluation, for $1 \leq i, j \leq m$.

3. Results

Figure 5 displays boxplots depicting the experimental results of the kappa index obtained from the validation process for different Random Forest configurations. Specifically, the boxplots illustrate the impact of varying the number of trees

within the range of 100 to 1000. Notably, an evident trend emerges, indicating that an increase in the number of trees leads to higher kappa values with reduced variability.

Comprehensive results for the evaluation metrics, including kappa, accuracy, sensitivity, specificity, and AUC, across all investigated configurations during the validation process, are presented in Table 6. These metrics provide a comprehensive assessment of the classification performance for each configuration, facilitating a comparative analysis and aiding in the identification of the most effective Random Forest setup.

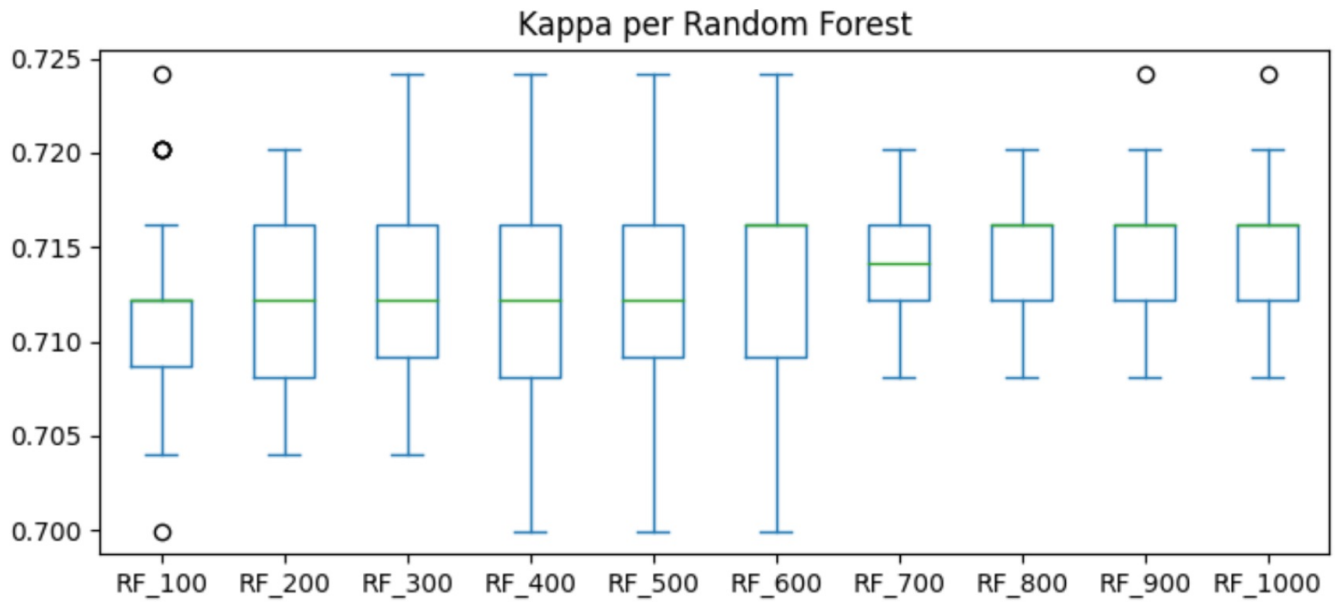

Figure 5. Boxplots of the kappa index results per Random Forest setup during the validation process. The number of trees was varied from 100 to 1000 trees. It is possible to see that the increase in the number of trees produces higher kappas with less variation.

Validation results

Classifier	Sensitivity		Specificity		AUC		Accuracy (%)		Kappa	
	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.	Mean	Std. Dev.
RF_100	0.9998	0.0006	0.5475	0.1086	0.9707	0.0172	98.18	0.44	0.6909	0.0931
RF_200	0.9998	0.0006	0.5478	0.1105	0.9775	0.0127	98.18	0.46	0.6911	0.0952
RF_300	0.9998	0.0006	0.5488	0.109	0.9797	0.0109	98.19	0.44	0.6919	0.0937
RF_400	0.9998	0.0006	0.5494	0.1098	0.9809	0.01	98.19	0.44	0.6924	0.0944
RF_500	0.9998	0.0006	0.5491	0.109	0.9816	0.0095	98.19	0.44	0.6923	0.0939
RF_600	0.9998	0.0006	0.5491	0.1094	0.982	0.0095	98.19	0.44	0.6922	0.0941
RF_700	0.9998	0.0006	0.5502	0.1093	0.9824	0.0092	98.19	0.44	0.6931	0.0935
RF_800	0.9998	0.0006	0.5514	0.1078	0.9826	0.0089	98.2	0.43	0.6943	0.0924
RF_900	0.9998	0.0006	0.551	0.1075	0.9827	0.0089	98.2	0.43	0.694	0.0921
RF_1000	0.9998	0.0006	0.5512	0.108	0.9829	0.0088	98.2	0.43	0.6941	0.0924

Table 6. Validation results of accuracy, kappa, sensitivity, specificity, and AUC, considering the sample mean and standard deviation, for all Random Forest configurations, varying the number of trees from 100 to 1000

From the analysis of the results of the kappa index in Table 6, it is possible to perceive that, considering only the sample mean and standard deviation, it is not possible to establish significant differences between the Random Forest configurations used. If we consider only two decimal places, the results are practically the same. However, observing the

behavior from the boxplots in Figure 5, we can see that the configurations with 800, 900 and 1000 trees have a higher median and show less variation in the results. In this way, we chose the intermediate configuration, with 900 trees, as the most adequate. Subsequently, we evaluated the performance of the chosen RF_900 configuration on the test dataset. A comprehensive summary of the test results can be found in Table 7.

RF_900					
Class	**Sensitivity**	**Specificity**	**AUC**	**Accuracy (%)**	**Kappa**
No docking	0.999	0.529	0.966	98.059	0.6744
Docking	0.529	0.999	0.966		
Weighted average	0.981	0.548	0.966		

Confusion matrix		**Predicted class**	
		No docking	Docking
True class	No docking	1236	1
	Docking	24	27

Table 7. Test results of accuracy, kappa, sensitivity, specificity, and AUC, for the 900-tree Random Forest

Table 8 presents the comprehensive test results obtained from the ensemble of 25 Random Forests, each comprising 900 trees and associated with a specific partition of the training set. The analysis reveals notable performance metrics for the classification task at hand. Sensitivity, reflecting the ability to correctly identify positive instances, exhibited a range of 91% to 96%, with a mean of 94%, a standard deviation of 1%, and a median of 94%. In terms of specificity, which gauges the accurate identification of negative instances, the mean value approached 78%, accompanied by a sample standard deviation of 2%, a median of 77%, and a range between 73% and 83%. The AUC value ranged from 92% to 94%, attaining a mean and median value of 93% and demonstrating a standard deviation of less than 1%. By employing the ensemble approach, the mean and median accuracy reached 94%, with a range of 91% to 96%. Furthermore, the mean and median kappa values, indicative of inter-rater agreement, were calculated at 0.47, accompanied by a standard deviation of 0.04, and ranged between 0.38 and 0.56. These findings underscore the effectiveness of the ensemble method in enhancing specificity while preserving high AUC and accuracy levels. Notably, AUC emerges as the most crucial metric due to its robustness in handling class imbalance and its ability to amalgamate sensitivity and specificity.

RF_900 Ensemble					
Training set	**Sensitivity**	**Specificity**	**AUC**	**Accuracy (%)**	**Kappa**
1	0.953	0.791	0.938	95.264	0.5443
2	0.941	0.753	0.923	94.0994	0.4722
3	0.936	0.828	0.934	93.6335	0.4774
4	0.93	0.734	0.923	93.0124	0.4195
5	0.937	0.772	0.933	93.7112	0.4616
6	0.932	0.79	0.931	93.1677	0.4456
7	0.946	0.753	0.931	94.6429	0.4984
8	0.938	0.753	0.935	93.7888	0.4583
9	0.93	0.819	0.94	93.0124	0.4459
10	0.943	0.753	0.926	94.3323	0.4831
11	0.937	0.791	0.942	93.7112	0.4681
12	0.956	0.773	0.928	95.5745	0.5559
13	0.935	0.753	0.919	93.4783	0.4451
14	0.935	0.791	0.94	93.4783	0.4582
15	0.932	0.808	0.935	93.1677	0.452
16	0.943	0.753	0.938	94.2547	0.4795
17	0.946	0.753	0.921	94.6429	0.4984
18	0.925	0.771	0.922	92.5466	0.4156
19	0.944	0.772	0.936	94.4099	0.4935
20	0.933	0.79	0.931	93.323	0.4519
21	0.947	0.753	0.929	94.7205	0.5023
22	0.91	0.808	0.935	90.9938	0.3775
23	0.943	0.772	0.936	94.3323	0.4898
24	0.939	0.772	0.935	93.8665	0.4684
25	0.931	0.79	0.925	93.0901	0.4425
Average	0.9377	0.7758	0.9314	93.7702	0.4682
Std. Dev.	0.0094	0.0241	0.0066	0.9409	0.0378
Median	0.9370	0.7720	0.9330	93.7112	0.4681
Min.	0.9100	0.7340	0.9190	90.9938	0.3775
Max.	0.9560	0.8280	0.9420	95.5745	0.5559

Confusion matrix		**Predicted class**	
		No docking	Docking
True class	No docking	1160	77
	Docking	11	40

Table 8. Test results of accuracy, kappa, sensitivity, specificity, and AUC, for the ensemble of 900-tree Random Forest

4. Discussion

The successful classification outcomes achieved through the application of Random Forests provide compelling evidence for the effectiveness of the pseudo-convolutional feature extraction method proposed in this study, pertaining to the identification of complex transcribed RNA sequences with superior sensitivity, specificity, and AUC. Notably, the Random Forest model with 900 trees exhibited the most optimal performance across the entirety of the test dataset.

Upon evaluating the parameters employed in the proposed sequence-based feature extraction technique, it is observed that utilizing a window size equal to 50% of the sequence length, a step size equivalent to 50% of the window size, and incorporating two hidden layers prove to be satisfactory in generating representative characteristics.

The Random Forest test utilizing 900 trees yielded notable outcomes, demonstrating the attainment of a high accuracy level and a considerable kappa performance of 98% and 0.67, respectively (refer to Table 7). Although the sensitivity for identifying non-interacting pairs exhibited a significantly high value of approximately 99.9%, the sensitivity for detecting pairs of proteins amenable to docking was only 52.9%, slightly surpassing the baseline of 50%. Consequently, the weighted sensitivity amounted to 98.1%, while the weighted specificity remained at a modest level of 54.8%. Despite the considerably high AUC value of 96.6%, this outcome indicates a substantial influence of class imbalance on the estimator, given that the population of undocked protein pairs is approximately 25 times greater than that of docked pairs.

To address this concern, a specific classification architecture was devised employing an ensemble approach with 25 Random Forest classifiers consisting of 900 trees each. Each classifier received a balanced fraction of the training set, comprising a 1/25 fraction of undocked pair instances, with instances from the docking class being repeated in every training subset. The results of this approach, as presented in Table 8, effectively improved specificity while maintaining a comparable level of sensitivity: the mean sensitivity amounted to 94%, whereas the mean specificity reached 83%, and the mean AUC achieved 93%. Furthermore, an analysis of individual classifier performance within the ensemble revealed their ability to generate favorable docking prediction outcomes, ultimately contributing to the overall result characterized by high accuracy, sensitivity, and AUC values, along with reasonably satisfactory specificity values.

5. Conclusion

This study endeavors to present a novel classifier architecture that leverages pseudo-convolutions and Random Forests to computationally estimate the molecular docking between two proteins. Our approach solely utilizes the sequence of characters representing the bases of each protein as input representation. These sequences undergo a pseudo-convolutional process, where co-occurrence matrices are calculated at each layer and concatenated into a comprehensive feature vector. The proposed architecture incorporates two layers of pseudo-convolutions. To assess its efficacy, we conducted validation using two widely recognized protein pair databases for evaluating digital molecular docking methods, namely Affinity Benchmark 3 and Negatome 2[2].

The output layer of the classification model was subjected to analysis, exploring the performance of a single Random Forest classifier and an ensemble consisting of 25 Random Forests, each comprising 900 trees. Our experimental results revealed that the hybrid architecture, comprising two layers of pseudo-convolution followed by a decision layer with an

ensemble of 25 Random Forests, yielded substantial improvements in accuracy, AUC, sensitivity, and specificity. Notably, this exploratory study, which focuses on classification rather than regression, demonstrates the feasibility of achieving robust estimates of molecular docking without relying on 3D molecule modeling. This finding indicates promising prospects for swiftly obtaining affinity estimates between proteins, offering valuable insights into drug interactions, molecular dynamics, and facilitating the intelligent design of pharmaceuticals, including vaccines and antiviral drugs. Such advancements are pivotal in reducing costs associated with laboratory analysis and expediting the delivery of solutions to the market and the broader society, particularly during epidemic outbreaks, pandemics, and the pressing need for additional vaccines.

As future directions, we aim to investigate the impact of augmenting the number of pseudo-convolution layers on the sensitivity and specificity of the method. Additionally, we plan to develop a regression-based variant to enable comparisons with state-of-the-art approaches based on 3D molecule modeling. Furthermore, we intend to explore the potential benefits of integrating the proposed architecture with other cutting-edge methods, with the overarching objective of achieving even higher levels of accuracy in predicting molecular docking outcomes.

Statements and Declarations

Acknowledgements

The authors thank the Brazilian agency Conselho Nacional de Desenvolvimento Científico e Tecnológico, CNPq, and Instituto do Complexo Econômico-Industrial da Saúde, iCEIS, Brazil, for the partial financial support for this research.

Conflicts of Interest

All authors declare they have no conflicts of interest.

References

1. [a, b, c] *Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. (2008). "Protein-protein docking benchmark version 3.0." Proteins. 73(3):705–709.*
2. [a, b, c, d] *Blohm P, Frishman G, Smialowski P, Goebels F, Wachinger B, et al. (2013). "Negatome 2.0: A database of non-interacting proteins derived by literature mining, manual annotation and protein structure analysis." Nucleic Acids Res. 42(Database issue):D396–400.*
3. [^] *Prisilla Jayanthi, Muralikrishna Iyyanki, Aruna Mothkuri, Prakruthi Vadakattu. (2020). Fourth industrial revolution: An impact on health care industry. In: Advances in artificial intelligence, software and systems engineering: Proceedings of the AHFE 2019 international conference on human factors in artificial intelligence and social computing, the AHFE international conference on human factors, software, service and systems engineering, and the AHFE international*

conference of human factors in energy, july 24-28, 2019, washington DC, USA 10.: Springer pp. 58–69.

4. ^*Jatinder Bali, Renu T. Bali. (2020). India and the fourth industrial revolution: How we should approach artificial intelligence in healthcare and biomedical research? The Journal of the Association of Physicians of India. 68(3):72–74.*
5. ^*Ki-Bong Kim, Kun-Hee Han. (2020). A study of the digital healthcare industry in the fourth industrial revolution. Journal of Convergence for Information Technology. 10(3):7–15.*
6. ^*João António Gomes de Melo e Castro e Melo, Nuno Miguel Faria Araújo. (2020). Impact of the fourth industrial revolution on the health sector: A qualitative study. Healthcare informatics research. 26(4):328–334.*
7. ^*Puneeta Ajmera, Vineet Jain. (2019). Modelling the barriers of health 4.0 – the fourth healthcare industrial revolution in india by TISM. Operations Management Research. 12(3-4):129–145.*
8. ^*Xiaobai Xiong. (2021). Bring technology home and stay healthy: The role of fourth industrial revolution and technology in improving the efficacy of health care spending. Technological Forecasting and Social Change. 165:120556.*
9. ^*Zhibo Pang, Heng Yuan, Yuan-Ting Zhang, Muthukumaran Packirisamy. (2018). Guest editorial health engineering driven by the industry 4.0 for aging society. IEEE Journal of Biomedical and Health Informatics. 22(6):1709–1710.*
10. ^*Safia Mahomed. (2018). Healthcare, artificial intelligence and the fourth industrial revolution: Ethical, social and legal considerations. South African Journal of Bioethics and Law. 11(2):93–95.*
11. ^*Antonio Celesti, Oliver Amft, Massimo Villari. Guest editorial special section on cloud computing, edge computing, internet of things, and big data analytics applications for healthcare industry 4.0. IEEE Transactions on Industrial Informatics.: IEEE 2019. pp. 454–456.*
12. ^*Amandeep Kaur, Ruchi Garg, Poonam Gupta. (2021). Challenges facing AI and big data for resource-poor healthcare system. In: 2021 second international conference on electronics and sustainable communication systems (ICESC).: IEEE pp. 1426–1433.*
13. ^*Vaibhav Thakare, Gauri Khire, Manisha Kumbhar. (2022). Artificial intelligence (AI) and internet of things (IoT) in healthcare: Opportunities and challenges. ECS Transactions. 107(1):7941.*
14. ^*Sarah Shafqat, Saira Kishwer, Raihan Ur Rasool, Junaid Qadir, Tehmina Amjad, et al. (2020). Big data analytics enhanced healthcare systems: A review. The Journal of Supercomputing. 76:1754–1799.*
15. ^*Cristina Elena Turcu, Cornel Octavian Turcu. (2013). Internet of things as key enabler for sustainable healthcare delivery. Procedia-Social and Behavioral Sciences. 73:251–256.*
16. ^*Ian K. Poyner, R. Simon Sherratt. (2019). Improving access to healthcare in rural communities—IoT as part of the solution. In: 3rd IET international conference on technologies for active and assisted living (TechAAL 2019).: IET pp. 1–6.*
17. ^*Amit Banerjee, Chinmay Chakraborty, Anand Kumar, Debabrata Biswas. (2020). Emerging trends in IoT and big data analytics for biomedical and health care technologies. Handbook of data science approaches for biomedical engineering. :121–152.*
18. ^*Brian A. Aguado, Joseph C. Grim, Adrianne M. Rosales, Jana J. Watson-Capps, Kristi S. Anseth. (2018). Engineering precision biomaterials for personalized medicine. Science translational medicine. 10(424):eaam8645.*
19. ^*Elisabetta Primiceri, Maria Serena Chiriacò, Francesca M. Notarangelo, Antonio Crocamo, Diego Ardissino, et al.*

(2018). Key enabling technologies for point-of-care diagnostics. Sensors. 18(11):3607.

20. ^Udayan Ghosh, Shen Ning, Yuzhu Wang, Yong Lin Kong. (2018). Addressing unmet clinical needs with 3D printing technologies. Advanced healthcare materials. 7(17):1800417.
21. ^Atheer Awad, Sarah J. Trenfield, Thomas D. Pollard, Jun Jie Ong, Moe Elbadawi, et al. (2021). Connected healthcare: Improving patient care using digital health technologies. Advanced Drug Delivery Reviews. 178:113958.
22. ^Curtis Lowery. (2020). What is digital health and what do i need to know about it? Obstetrics and Gynecology Clinics. 47(2):215–225.
23. ^Maksut Senbekov, Timur Saliev, Zhanar Bukeyeva, Aigul Almabayeva, Marina Zhanaliyeva, et al. (2020). The recent progress and applications of digital technologies in healthcare: A review. International journal of telemedicine and applications. 2020.
24. ^Nic Fleming. (2018). How artificial intelligence is changing drug discovery. Nature. 557(7706):S55–S55.
25. ^Debleena Paul, Gaurav Sanap, Snehal Shenoy, Dnyaneshwar Kalyane, Kiran Kalia, et al. (2021). Artificial intelligence in drug discovery and development. Drug discovery today. 26(1):80.
26. ^Prashansa Agrawal. (2018). Artificial intelligence in drug discovery and development. Journal of Pharmacovigilance. 6(2):1000e173.
27. ^Veer Patel, Manan Shah. (2022). Artificial intelligence and machine learning in drug discovery and development. Intelligent Medicine. 2(3):134–140.
28. ^Alex Zhavoronkov. (2018). Artificial intelligence for drug discovery, biomarker development, and generation of novel chemistry. Molecular Pharmaceutics. 15(10):4311–4313.
29. ^José Jiménez-Luna, Francesca Grisoni, Nils Weskamp, Gisbert Schneider. (2021). Artificial intelligence in drug discovery: Recent advances and future perspectives. Expert opinion on drug discovery. 16(9):949–959.
30. ^Alex Zhavoronkov, Quentin Vanhaelen, Tudor I. Oprea. (2020). Will artificial intelligence for drug discovery impact clinical pharmacology? Clinical Pharmacology & Therapeutics. 107(4):780–785.
31. ^Jianyuan Deng, Zhibo Yang, Iwao Ojima, Dimitris Samaras, Fusheng Wang. (2022). Artificial intelligence in drug discovery: Applications and techniques. Briefings in Bioinformatics. 23(1).
32. a, bLeonardo G. Ferreira, Ricardo N. dos Santos, Glaucius Oliva, Adriano D. Andricopulo. (2015). Molecular docking and structure-based drug design strategies. Molecules. 20(7):13384–13421.
33. a, bAgnihotry S, Pathak RK, Srivastav A, Shukla PK, Gautam B. (2020). "Molecular docking and structure-based drug design". Computer-aided drug design. :115–131.
34. a, b, c, dSaikia S, Bordoloi M. (2019). "Molecular docking: Challenges, advances and its use in drug discovery perspective". Current drug targets. 20(5):501–521.
35. a, bRaval K, Ganatra T. (2022). "Basics, types and applications of molecular docking: A review". IP International Journal of Comprehensive and Advanced Pharmacology. 7(1):12–16.
36. a, bOtari KV, Menkudale AC, Kulkarni VC, Galave VB, Khemnar MD. (2021). "A review on molecular docking". International Research Journal of Pure and Applied Chemistry. 22(3):60–68.

37. [a, b, c]*Pinzi L, Rastelli G. (2019). "Molecular docking: Shifting paradigms in drug discovery". International journal of molecular sciences. 20(18):4331.*

38. [a, b]*De Paris R, Ruiz DA, De Souza ON. (2015). "A cloud-based workflow approach for optimizing molecular docking simulations of fully-flexible receptor models and multiple ligands". In: 2015 IEEE 7th international conference on cloud computing technology and science (CloudCom).: IEEE pp. 495–498.*

39. [a, b]*Quevedo CV, De Paris R, Ruiz DD, de Souza ON. (2014). "A strategic solution to optimize molecular docking simulations using fully-flexible receptor models". Expert Systems with Applications. 41(16):7608–7620.*

40. [a, b]*Xu J, Li J, Cai Y. (2017). "Molecular docking simulation based on CPU-GPU heterogeneous computing". In: Advanced parallel processing technologies: 12th international symposium, APPT 2017, santiago de compostela, spain, august 29, 2017, proceedings 12.: Springer pp. 27–37.*

41. [a, b]*Altuntaş S, Bozkus Z, Fraguela BB. (2016). "GPU accelerated molecular docking simulation with genetic algorithms". In: Applications of evolutionary computation: 19th european conference, EvoApplications 2016, porto, portugal, march 30–april 1, 2016, proceedings, part II 19.: Springer pp. 134–146.*

42. [^]*Yuriev E, Agostino M, Ramsland PA. (2011). "Challenges and advances in computational docking: 2009 in review". Journal of Molecular Recognition. 24(2):149–164.*

43. [^]*Alonso H, Bliznyuk AA, Gready JE. (2006). "Combining docking and molecular dynamic simulations in drug design". Medicinal research reviews. 26(5):531–568.*

44. [^]*Bello M, Martínez-Archundia M, Correa-Basurto J. (2013). "Automated docking for novel drug discovery". Expert opinion on drug discovery. 8(7):821–834.*

45. [a, b]*Salmaso V, Moro S. (2018). "Bridging molecular docking to molecular dynamics in exploring ligand-protein recognition process: An overview". Frontiers in pharmacology. 9:923.*

46. [^]*Kitchen DB, Decornez H, Furr JR, Bajorath J. (2004). "Docking and scoring in virtual screening for drug discovery: Methods and applications". Nature reviews Drug discovery. 3(11):935–949.*

47. [^]*Okimoto N, Futatsugi N, Fuji H, Suenaga A, Morimoto G, et al. (2009). "High-performance drug discovery: Computational screening by combining docking and molecular dynamics simulations". PLoS computational biology. 5(10):e1000528.*

48. [^]*Taylor RD, Jewsbury PJ, Essex JW. (2002). "A review of protein-small molecule docking methods". Journal of computer-aided molecular design. 16:151–166.*

49. [a, b]*Di Biasi L, Fino R, Parisi R, Sessa L, Cattaneo G, et al. (2016). "Novel algorithm for efficient distribution of molecular docking calculations". In: Advances in artificial life, evolutionary computation and systems chemistry: 10th italian workshop, WIVACE 2015, bari, italy, september 22-25, 2015, revised selected papers 10.: Springer pp. 65–74.*

50. [a, b]*Dong D, Xu Z, Zhong W, Peng S. (2018). "Parallelization of molecular docking: A review". Current Topics in Medicinal Chemistry. 18(12):1015–1028.*

51. [^]*Hecht D, Fogel GB. (2009). "Computational intelligence methods for docking scores". Current Computer-Aided Drug Design. 5(1):56–68.*

52. ^*Ballester PJ, Mitchell JB. (2010). "A machine learning approach to predicting protein–ligand binding affinity with applications to molecular docking". Bioinformatics. 26(9):1169–1175.*
53. ^*Yang C, Chen EA, Zhang Y. (2022). "Protein–ligand docking in the machine-learning era". Molecules. 27(14):4568.*
54. ^*Ashtawy HM, Mahapatra NR. (2014). "Molecular docking for drug discovery: Machine-learning approaches for native pose prediction of protein-ligand complexes". In: Computational intelligence methods for bioinformatics and biostatistics: 10th international meeting, CIBB 2013, nice, france, june 20-22, 2013, revised selected papers 10.: Springer pp. 15–32.*
55. ^*Alghamedy F, Bopaiah J, Jones D, Zhang X, Weiss HL, et al. (2018). "Incorporating protein dynamics through ensemble docking in machine learning models to predict drug binding". AMIA Summits on Translational Science Proceedings. 2018:26.*
56. ^*Hsin KY, Ghosh S, Kitano H. (2013). "Combining machine learning systems and multiple docking simulation packages to improve docking prediction reliability for network pharmacology". PloS one. 8(12):e83922.*
57. ^*Terayama K, Iwata H, Araki M, Okuno Y, Tsuda K. (2018). "Machine learning accelerates MD-based binding pose prediction between ligands and proteins". Bioinformatics. 34(5):770–778.*
58. a, b*Ma D, Guo Y, Luo J, Pu X, Li M. (2014). "Prediction of protein–protein binding affinity using diverse protein–protein interface features". Chemometrics and Intelligent Laboratory Systems. 138:7–13.*
59. a, b*Veit-Acosta M, de Azevedo Junior WF. (2021). "The impact of crystallographic data for the development of machine learning models to predict protein-ligand binding affinity". Current Medicinal Chemistry. 28(34):7006–7022.*
60. a, b*Li XL, Zhu M, Li XL, Wang HQ, Wang S. (2012). "Protein-protein interaction affinity prediction based on interface descriptors and machine learning". In: Intelligent Computing Theories and Applications: 8th International Conference, ICIC 2012, Huangshan, China, July 25-29, 2012 Proceedings 8.: Springer pp. 205–212.*
61. a, b*Kanakala GC, Aggarwal R, Nayar D, Priyakumar UD. (2023). "Latent biases in machine learning models for predicting binding affinities using popular data sets". ACS Omega.*
62. a, b*Bitencourt-Ferreira G, de Azevedo WF. (2019). "Machine learning to predict binding affinity". Docking Screens for Drug Discovery. :251–273.*
63. a, b*Druchok M, Yarish D, Garkot S, Nikolaienko T, Gurbych O (2021). "Ensembling machine learning models to boost molecular affinity prediction." Computational Biology and Chemistry. 93:107529.*
64. a, b*Heck GS, Pintro VO, Pereira RR, Levin NMB, de Azevedo WF (2017). "Supervised machine learning methods applied to predict ligand-binding affinity." Current Medicinal Chemistry. 24(23):2459–2470.*
65. a, b, c*Jiménez-Luna J, Cuzzolin A, Bolcato G, Sturlese M, Moro S (2020). "A deep-learning approach toward rational molecular docking protocol selection." Molecules. 25(11):2487.*
66. a, b, c*Gentile F, Agrawal V, Hsing M, Ton AT, Ban F, et al. (2020). "Deep docking: A deep learning platform for augmentation of structure based drug discovery." ACS Central Science. 6(6):939–949.*
67. a, b, c*Morrone JA, Weber JK, Huynh T, Luo H, Cornell WD (2020). "Combining docking pose rank and structure with deep learning improves protein–ligand binding mode prediction over a baseline docking approach." Journal of*

Chemical Information and Modeling. 60(9):4170–4179.

68. [a, b, c] *Yang L, Yang G, Chen X, Yang Q, Yao X, et al. (2021). "Deep scoring neural network replacing the scoring function components to improve the performance of structure-based molecular docking." ACS Chemical Neuroscience. 12(12):2133–2142.*
69. [^] *de Lima SML, da Silva-Filho AG, dos Santos WP (2016). "Detection and classification of masses in mammographic images in a multi-kernel approach." Computer Methods and Programs in Biomedicine. 134:11–29.*
70. [^] *de Santana MA, Silva Pereira JM, da Silva FL, de Lima NM, de Sousa FN, et al. (2018). "Breast cancer diagnosis based on mammary thermography and extreme learning machines." Research on Biomedical Engineering. 34:45–53.*
71. [^] *Gomes JC, de Freitas Barbosa VA, de Santana MA, Bandeira J, Valença MJS, et al. (2020). "IKONOS: An intelligent tool to support diagnosis of Covid-19 by texture analysis of x-ray images." Research on Biomedical Engineering. :1–14.*
72. [^] *de Freitas Barbosa VA, Gomes JC, de Santana MA, de Almeida JE, de Souza RG, et al. (2021). "Heg.IA: An intelligent system to support diagnosis of Covid-19 based on blood tests." Research on Biomedical Engineering. :1–18.*
73. [^] *de Santana MA, Silva Pereira JM, da Silva FL, de Lima NM, de Sousa FN, et al. (2018). "Breast cancer diagnosis based on mammary thermography and extreme learning machines." Research on Biomedical Engineering. 34(1):45–53.*
74. [^] *Espinola CW, Gomes JC, Silva Pereira JM, dos Santos WP (2021). "Detection of major depressive disorder using vocal acoustic analysis and machine learning—an exploratory study." Research on Biomedical Engineering. 37(1):53–64.*
75. [^] *Espinola CW, Gomes JC, Silva Pereira JM, dos Santos WP (2021). "Vocal acoustic analysis and machine learning for the identification of schizophrenia." Research on Biomedical Engineering. 37(1):33–46.*
76. [a, b] *Cordeiro FR, Santos WP, Silva-Filho AG (2016). "A semi-supervised fuzzy GrowCut algorithm to segment and classify regions of interest of mammographic images." Expert Systems with Applications. 65:116–126.*
77. [^] *Azevedo WW, Lima SML, Fernandes IMM, Rocha ADD, Cordeiro FR, et al. (2015). "Fuzzy morphological extreme learning machines to detect and classify masses in mammograms." In: 2015 IEEE International Conference on Fuzzy Systems (fuzz-IEEE).: IEEE pp. 1–8.*
78. [^] *De Oliveira APS, de Santana MA, Andrade MKS, Gomes JC, Rodrigues MCA, et al. (2020). "Early diagnosis of parkinson's disease using EEG, machine learning and partial directed coherence." Research on Biomedical Engineering. 36:311–331.*
79. [^] *Gomes JC, de Santana MA, Masood AI, de Lima CL, dos Santos WP (2023). "COVID-19's influence on cardiac function: A machine learning perspective on ECG analysis." Medical & Biological Engineering & Computing. :1–25.*
80. [^] *de Santana MA, de Freitas Barbosa VA, de Lima RCF, dos Santos WP (2022). "Combining deep-wavelet neural networks and support-vector machines to classify breast lesions in thermography images." Health and Technology. :1–13.*
81. [^] *Shirahige L, Leimig B, Baltar A, Bezerra A, de Brito CVF, et al. (2022). "Classification of parkinson's disease motor*

phenotype: A machine learning approach." Journal of Neural Transmission. :1–15.

82. ^*Fonseca FS, Torcate AS, Da Silva ACG, Freire VHW, De Farias GPM, et al. (2022). "Early prediction of generalized infection in intensive care units from clinical data: A committee-based machine learning approach." In: 2022 IEEE Latin American Conference on Computational Intelligence (LA-CCI).: IEEE pp. 1–6.*
83. ^*de Santana MA, dos Santos WP (2022). "A deep-wavelet neural network to detect and classify lesions in mammographic images." Research on Biomedical Engineering. 38(4):1051–1066.*
84. ^*de Freitas Barbosa VA, da Silva AF, de Santana MA, de Azevedo RR, de Lima RCF, et al. (2022). "Deep-wavelets and convolutional neural networks to support breast cancer diagnosis on thermography images." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. :1–19.*
85. ^*Espinola CW, Gomes JC, Silva Pereira JM, dos Santos WP (2022). "Detection of major depressive disorder, bipolar disorder, schizophrenia and generalized anxiety disorder using vocal acoustic analysis and machine learning: An exploratory study." Research on Biomedical Engineering. 38(3):813–829.*
86. ^*Gomes JC, Rodrigues MCA, dos Santos WP (2022). "ASTERI: Image-based representation of EEG signals for motor imagery classification." Research on Biomedical Engineering. 38(2):661–681.*
87. ^*de Freitas Barbosa VA, Gomes JC, de Santana MA, de Lima CL, Calado RB, et al. (2022). "Covid-19 rapid test by combining a random forest-based web system and blood tests." Journal of Biomolecular Structure and Dynamics. 40(22):11948–11967.*
88. ^*de Souza RG, Lucas e Silva GS, dos Santos WP, de Lima ME, Alzheimer's Disease Neuroimaging Initiative (2021). "Computer-aided diagnosis of alzheimer's disease by MRI analysis and evolutionary computing." Research on Biomedical Engineering. 37:455–483.*
89. ^[a, b]*Fan FJ, Shi Y (2022). "Effects of data quality and quantity on deep learning for protein-ligand binding affinity prediction." Bioorganic & Medicinal Chemistry. 72:117003.*
90. ^[a, b]*Ahmed A, Mam B, Sowdhamini R (2021). "DEELIG: A deep learning approach to predict protein-ligand binding affinity." Bioinformatics and Biology Insights. 15:11779322211030364.*
91. ^[a, b]*Stepniewska-Dziubinska MM, Zielenkiewicz P, Siedlecki P (2018). "Development and evaluation of a deep learning model for protein–ligand binding affinity prediction." Bioinformatics. 34(21):3666–3674.*
92. ^[a, b]*Gomes JC, Masood AI, de S. Silva LH, da Cruz Ferreira JRB, Freire Junior AA, et al. (2021). "Covid-19 diagnosis by combining RT-PCR and pseudo-convolutional machines to characterize virus sequences." Scientific Reports. 11(1):1–28.*
93. ^[a, b]*Crampon K, Giorkallos A, Deldossi M, Baud S, Steffenel LA. (2022). "Machine-learning methods for ligandprotein molecular docking." Drug Discovery Today. 27(1):151–164.*
94. ^[a, b]*Chandak T, Mayginnes JP, Mayes H, Wong CF. (2020). "Using machine learning to improve ensemble docking for drug discovery." Proteins: Structure, Function, and Bioinformatics. 88(10):1263–1270.*
95. ^*Kashyap J, Datta D. (2022). "Drug repurposing for SARS-CoV-2: A high-throughput molecular docking, molecular dynamics, machine learning, and DFT study." Journal of Materials Science. 57(23):10780–10802.*
96. ^*Li C, Hua Y, Pan D, Qi L, Xiao C, et al. (2023). "A rapid selection strategy for umami peptide screening based on*

machine learning and molecular docking." Food Chemistry. 404:134562.

97. ^*de Oliveira TA, Medaglia LR, Maia EHB, Assis LC, de Carvalho PB, et al. (2022). "Evaluation of docking machine learning and molecular dynamics methodologies for DNA-ligand systems." Pharmaceuticals. 15(2):132.*
98. ^*Yang S, Li S, Chang J. (2022). "Discovery of cobimetinib as a novel a-FABP inhibitor using machine learning and molecular docking-based virtual screening." RSC Advances. 12(21):13500–13510.*
99. ^*Choi J, Lee J. (2021). "V-dock: Fast generation of novel drug-like molecules using machine-learning-based docking score and molecular optimization." International Journal of Molecular Sciences. 22(21):11635.*
100. ^*Ricci-Lopez J, Aguila SA, Gilson MK, Brizuela CA. (2021). "Improving structure-based virtual screening with ensemble docking and machine learning." Journal of Chemical Information and Modeling. 61(11):5362–5376.*
101. ^*Nogueira MS, Koch O. (2019). "The development of target-specific machine learning models as scoring functions for docking-based target prediction." Journal of Chemical Information and Modeling. 59(3):1238–1252.*
102. ^*Lin X, Li X, Lin X. (2020). "A review on applications of computational methods in drug screening and design." Molecules. 25(6):1375.*
103. ^*Wu F, Zhou Y, Li L, Shen X, Chen G, et al. (2020). "Computational approaches in preclinical studies on drug discovery and development." Frontiers in Chemistry. 8:726.*
104. ^*Zhao J, Cao Y, Zhang L. (2020). "Exploring the computational methods for protein-ligand binding site prediction." Computational and Structural Biotechnology Journal. 18:417–426.*
105. ^*Al-Khafaji K, Al-Duhaidahawi D, Tok TT. (2021). "Using integrated computational approaches to identify safe and rapid treatment for SARS-CoV-2." Journal of Biomolecular Structure and Dynamics. 39(9):3387–3395.*
106. ^*Aftab SO, Ghouri MZ, Masood MU, Haider Z, Khan Z, et al. (2020). "Analysis of SARS-CoV-2 RNA-dependent RNA polymerase as a potential therapeutic drug target using a computational approach." Journal of Translational Medicine. 18(1):1–15.*
107. ^*Cui W, Aouidate A, Wang S, Yu Q, Li Y, et al. (2020). "Discovering anti-cancer drugs via computational methods." Frontiers in Pharmacology. 11:733.*
108. ^*Norman RA, Ambrosetti F, Bonvin AMJJ, Colwell LJ, Kelm S, et al. (2020). "Computational approaches to therapeutic antibody design: Established methods and emerging trends." Briefings in Bioinformatics. 21(5):1549–1567.*
109. ^*Wang X, Terashi G, Christoffer CW, Zhu M, Kihara D. (2020). "Protein docking model evaluation by 3D deep convolutional neural networks." Bioinformatics. 36(7):2113–2118.*
110. ^*Parvez MSA, Karim MA, Hasan M, Jaman J, Karim Z, et al. (2020). "Prediction of potential inhibitors for RNA-dependent RNA polymerase of SARS-CoV-2 using comprehensive drug repurposing and molecular docking approach." International Journal of Biological Macromolecules. 163:1787–1797.*
111. ^*Wu C, Liu Y, Yang Y, Zhang P, Zhong W, et al. (2020). "Analysis of therapeutic targets for SARS-CoV-2 and discovery of potential drugs by computational methods." Acta Pharmaceutica Sinica B. 10(5):766–788.*
112. ^*Breiman L. (2001). "Random forests." Machine learning. 45:5–32.*
113. ^*Cutler A, Cutler DR, Stevens JR. (2012). "Random forests." In: Ensemble machine learning: Methods and*

applications. :157–175.

114. ^*Goel E, Abhilasha E, Goel E, Abhilasha E. (2017). "Random forest: A review." International Journal of Advanced Research in Computer Science and Software Engineering. 7(1):251–257.*
115. a, b*Duda R, Hart P, Stork DG. (2001). "Pattern classification." John Wiley; Sons.*
116. ^*Cordeiro FR, dos Santos WP, Silva-Filho AG. (2017). "Analysis of supervised and semi-supervised GrowCut applied to segmentation of masses in mammography images." Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization. 5(4):297–315.*
117. ^*Lima S, Azevedo W, Cordeiro F, Silva-Filho A, Santos W. (2015). "Feature extraction employing fuzzy-morphological decomposition for detection and classification of mass on mammograms." In: Conference proceedings: Annual international conference of the IEEE engineering in medicine and biology society IEEE engineering in medicine and biology society Annual conference. pp. 801–804.*
118. ^*Cordeiro FR, Bezerra KFP, dos Santos WP. (2017). "Random walker with fuzzy initialization applied to segment masses in mammography images." In: 2017 IEEE 30th international symposium on computer-based medical systems (CBMS).: Thessaloniki pp. 156–161.*
119. ^*Cordeiro FR, Santos WP, Silva-Filhoa AG. (2013). "Segmentation of mammography by applying GrowCut for mass detection." Studies in health technology and informatics. 192:87.*
120. ^*Cordeiro FR, Santos WP, Silva-Filho AG. (2016). "An adaptive semi-supervised fuzzy GrowCut algorithm to segment masses of regions of interest of mammographic images." Applied Soft Computing. 46:613–628.*