JCBI
Journal of Computing & Biomedical Informatics Volume 04 Issue 01
ISSN: 2710 - 1606 2022

Research Article
https://doi.org/10.56979/401/2022/95

Neural Network-Based Prediction of Potential Ribonucleic Acid Aptamers to Target Protein

Muhammad Azhar Mahmood[1], Hassaan Malik[1,2*], Ali Haider Khan[2], Muhammad Adnan[1], and Muhammad Imran Ali Khan[1]

[1]Department of Computer Science, National College of Business Administration & Economics Lahore, Multan Sub Campus, Multan, 60000, Pakistan
[2]School of Systems & Technology, Department of Computer Science, University of Management and Technology, Lahore, Pakistan.
*Corresponding Author: Hassaan Malik. Email: f2019288004@gmail.com.

Received: September 28, 2022 **Accepted:** November 28, 2022 **Published:** December 29, 2022.

Abstract: Aptamers are short strands of nucleic acid with a single strand that may unite to target a certain molecule in a selective and specific manner. SELEX experiments are the typical method used for identifying aptamers in vitro (systematic evolution of ligands by exponential enrichment). Several different computational methods have been developed to locate aptamers. The purpose of this research is to identify and make predictions on the possible RNA aptamers that may be used to target the protein. To do this, we propose the use of a multi-layer perceptron neural network with sixteen layers that are trained to locate possible aptamers of a protein target. This network is trained by extracting the main properties of RNA sequences. The outcome of our proposed model is compared to the output of two well-known machine learning classifiers, namely random forest (RF) and support vector machine (SVM). Additionally, we undertake the independent testing of our model on the benchmark dataset, which allows us to reach the highest accuracy possible. As a consequence of this, our model obtains an accuracy of 98.44% and an MCC of 0.9123 during the 15-fold cross-validation, and it achieves an accuracy of 98.10% and an MCC of 0.9354 when the leave-one-out cross-validation is performed. We are certain that our approach will contribute to a reduction in the amount of money and time spent on in vitro testing. Therefore, restricting the length of the initial pool of potential nucleic acid pattern combinations.

Keywords: Neural Network; Aptamers; Ribonucleic Acid Aptamers.

.

1. Introduction

Andy Ellington was the one who first coined the word "Aptamers" [1]. They are single-stranded nucleic acids that are relatively short and include DNA or RNA sequences that combine to target the molecule in question [2]. Examples of such molecules include carbohydrates, toxins, peptides, and proteins. SELEX is an in vitro approach that is used to identify aptamers for the protein target from a large oligonucleotide library [3]. Beginning in the early 1990s, a wide range of aptamers were used in many applications to target the illness, such as in medical trials for the identification of various disorders [4]. Aptamers, in particular, have yielded large outputs in comparison to protein antibodies as a result of their easy chemical amalgamation, low immunogenicity, and thermal stability [5]. For instance, He et al. [6] established an innovative method by selecting the DNA aptamers, which identified drug-resistant ovarian cancer using the SELEX technique. Mateja et al. [7] produced a method based on SELEX cells for identifying the DNA aptamers that are used to find the existence of non-small lung carcinoma (NSLC) on the cell surface. Su et al. [8] designed a sensor containing the feature of detecting bisphenol A in a real-time environment by using the sequences of DNA aptamer. The process of the SELEX method consists of a variant type of initial steps

Mateja et al. [7] developed a technique that is based on SELEX cells for detecting the DNA aptamers that are utilized to determine whether or not non-small lung cancer (NSLC) is present on the cell surface. Using the sequences of DNA aptamer, the researchers Su et al. [8] developed a sensor that is capable of detecting bisphenol A in a real-time setting. A variety of preliminary procedures, including amplification, washing, binding, and incubation, are involved in the SELEX method's process [9]. The first step in the SELEX procedure is to create a library of single-stranded nucleic acid (S-SNA) sequences. These libraries typically have 1015 random S-SNA sequences in them, but only a select few sequences with a high affinity are reversed [10]. The whole SELEX procedure is comprised of around 15 rounds, and completing the assignment may take anything from a few days to several months [11]. An efficient and significant computational strategy that shortens the duration of the experimental phases while also lowering their associated costs [9] [12] is needed.

In addition to in vivo techniques, a great number of computational methodologies have been developed to determine the sequences of aptamers [10-12]. However, to the best of our knowledge, these methodologies are not yet being used to find novel aptamers for a target [13-14]. In addition, a small number of mathematical models are used in the process of selecting aptamers for the only purpose of an individual target [15-18]. In this study work, we created a computational system that is capable of producing possible RNA aptamers that target the protein. We also extract the dominant and important patterns employed for interacting with RNA and protein molecules from the protein-RNA (P-RNA) sequences. The multi-layer perceptron (MLP) classifier that has been presented has been developed with the help of the patterns that have been identified as coming from P-RNA complexes. The results of our newly developed MLP provide accurate predictions about the candidates for substantial prospective aptamers among the collection of strong aptamers. In addition, the contribution of this research may be summarised as follows::

1. To identify the prospective candidates for aptamers, we suggested using an MLP model. We also compare the results of our model MLP with those of two well-known machine learning classifiers such as RF and SVM in terms of sensitivity (SN), specificity (SP), accuracy (ACC), Matthews' correlation coefficient (MCC), positive predictive value (PPV), and negative predictive value (NPV) (NPV). The performance of the proposed model, RF, and SVM was validated by applying 15-fold cross-validation and leave-one-out validation.

2. We also applied our proposed classifiers to the publically available benchmark dataset designed by [13] and compared our results with it.

3. The result reveals that our proposed MLP classifier is more effective than the traditional SELEX process in finding the aptamers to target a protein.

The remaining portion of this study is structured as follows: Section 2 discusses the Literature review. Section 3 provides the Materials and method. In section 4, results and discussion are presented, and in the last section 5, this study is concluded.

2. Literature Review

The reviews of the relevant previous researchers are presented in this chapter. Several of the studies that have been reviewed in this chapter are examples of how the literature review can be used to find research gaps and identify appropriate methodologies. Some of the studies also expressed, that the collection of up-to-date knowledge about the linked research is used for, the research purpose that was supposed for the studies. The evaluation of the research literature extracting the key points for purposed research.

Ribonucleic acid-binding proteins (RBPs) for short, are proteins that bind to double- or single-stranded RNA in cells and take part in the creation of RNA Protein complexes. RBPs play a key function in controlling many things. However, it is still unclear how RBPs choose which subsequence target RNAs to search for and why they do so. Discovering the appropriate RNA transcription factor binding sites is a very crucial stage in the process of gaining a deeper understanding of the operation of many biological processes.

The RBPCNN model is a simple and effective deep-learning convolution neural network that integrates information about evolution with raw RNA sequences. The model is introduced in the publication [19], which also contains the RBPCNN model. Additionally, the automated extraction of the binding sequence motifs might assist them in gaining a better understanding of how RBPs bind to their respective targets. The findings of the trials indicate that RBPCNN performs much better than the approaches that

are considered to be the best at the moment. To be more exact, the average area under the receiver operator curve improved by 2.67 percent, while the average mean accuracy improved by 8.03 percent. In comparison to the most cutting-edge approaches, this integration enabled us to achieve very successful outcomes. They then created drawings of the motifs that the RBPCNN model instructed them to draw and compared those pictures to motifs that had already been discovered and published in the CISBP-RNA database. In addition to this, they created a graphic showing the standard deviation of the conservation scores that the deep learning kernels had acquired.

In a wide number of diagnostic and therapeutic applications, aptamers are strong contenders to monoclonal antibodies as the antibody of choice. To hasten the process of exponentially enriching ligands by SELEX (systematic evolution of ligands by exponential enrichment), in silico methods have been developed. SELEX is notorious for being a time-consuming and costly endeavor. During this study [20], in silico generation of aptamer sequences targeting CD13 was carried out using a genetic algorithm (GA) implementation that included a prediction model as part of its fitness function in two phases. This was accomplished by utilizing a genetic programming language. In the beginning, the purpose of the model was to make predictions about the RNA sequences of CD13. This was accomplished by using the sequence and structure of macromolecules derived from ribonucleic acid–protein complexes found on PDB. The model achieved an F1-score of 0.9273 and an overall accuracy of 92.72 percent on an independent data set by making use of the 196 characteristics that performed the best.

In the second step of the process, GA was used to generate new sequences by using the anticipated outcomes as the starting generation for the algorithm. [20], using GA, generated aptamers that had a higher GA score than their parent oligonucleotide sequences. This was accomplished by using GA. The findings of the docking and molecular dynamics simulations provide evidence that this strategy is successful. With the aid of this research, aptamers may be chosen according to a broad range of biochemical characteristics.

According to the findings of a study carried out by [21], the Singapore grouper iridovirus (SGIV) is responsible for causing significant economic losses in mariculture. There is a critical and immediate need for efficient therapies for SGIV infection. There is a rich variety of medicinal plant sources in China. Medicinal herbs have been used to treat a wide variety of illnesses for a long time and have significant therapeutic capabilities. The majority of the time, reverse-transcription quantitative real-time PCR is used to precisely diagnose a viral infection and evaluate the efficacy of a potential antiviral medication. However, their usefulness is restricted since the necessary processes and reagents are time-consuming and labor-intensive. Aptamers, which work by amplifying signals, have been included in specific biosensors to locate infections and illnesses with a high degree of precision. The purpose of this research was to develop an aptamer-based high-throughput screening (AHTS) approach that would facilitate the efficient selection and assessment of medicinal plant components about their effectiveness against SGIV infection. "aptamer-based high-throughput screening" is what "AHTS" stands for in the scientific community. The Q2-AHTS method, which has been classified as being sensitive, swift, and exact, is a speedy and effective strategy for selecting medicinal plant medications for the treatment of SGIV. The AHTS method not only cut down on the amount of time and money spent on experiments, but it also sped up the whole screening process for more effective compounds.

AHTS should be suitable for the speedy identification of components that are efficient against other viruses, according to [21]. [21] Non-coding RNAs (ncRNAs), which account for the majority of the genome, perform a variety of complex and specific activities, and it is essential to understand these roles to get an understanding of almost every aspect of cancer. This extensive group of chemicals is responsible for important activities in the control mechanisms of a variety of cellular processes. Regulatory mechanisms that are mediated by interactions between long noncoding RNAs (lncRNAs) and RNA-binding proteins have been associated with several different types of cancer.

Their effects are made possible by networks that regulate lncRNA and RBP stability, ncRNA Metabolism including N6-methyladenosine (m6A) and alternative splicing, subcellular localization, and a wide variety of other cancer-related pathway processes. This review [22] investigated the reciprocal interaction that exists between long noncoding RNAs (lncRNAs) and RBPs, as well as their participation in epigenetic regulation through histone modifications and their essential role in cancer treatment resistance. Other properties of RBPs, such as the structural domains they include, give further insight into how lncRNAs and RBPs interact with one another and how they carry out their separate biological functions. This is because structural domains are responsible for the organization of RBPs. In addition, the present state-of-

the-art information, which is made possible by machine learning and deep learning approaches, disentangles such linkages in more detail to further increase our comprehension of the subject matter. In addition, operations based on RNA are described in this article as a possible alternative therapy option that people afflicted with cancer would want to take into consideration. Because of the advancements that have been achieved in next-generation sequencing, several innovative approaches have been created. Among these techniques are the cross-linking and immunoprecipitation-seq (CLIP-seq) method, the RIP-Chip method, the RIP-Seq method, the MS2 trapping technique, and many more. These methods may be divided into two distinct categories: those that concentrate on RNA, and those that focus on proteins. A method that focuses on RNA makes an effort to identify every protein that can bind to an RNA of interest. When using a protein-centric technique, on the other hand, the objective is to find any RNAs that can bind to a certain protein of interest. This may be a challenge since there are so many different proteins. Molecular docking (MD) and machine learning (ML)-based methodologies are the two primary categories of LPI computational methods that are used. The majority of laboratories use MD-based techniques as their primary method for LPI prediction. The majority of MD-based programs, with a few notable exceptions such as HexServer, are both costly and time-consuming to operate.

In his study, [23] revealed an innovative deep learning methodology for predicting API that he dubbed AptaNet. AptaNet is one of a kind since it is capable of predicting API by using the sequence-based attributes of aptamers in addition to the physicochemical and conformational properties of targets. In addition to that, we use a deep neural network as well as a system for balancing things out. To determine how effectively AptaNet functions, [23] has conducted a great deal of research and testing. Experiments show that AptaNet has higher accuracy than other methods that were investigated for this study on their 32 benchmark datasets, where Aptamers were encoded utilizing two distinct strategies (k-mer frequency and reverse complement k-mer frequency). This was determined by comparing the results of AptaNet to those of the other methods. By using 24 physicochemical and structural features of the proteins, amino acid composition (AAC) and pseudo amino acid composition (PseAAC) were employed to represent target information. [23] used a neighborhood cleaning technique to solve the imbalance problem that was present in the data. The cornerstone for the building of the predictor was a deep neural network, and the random forest approach was utilized to determine which characteristics were the most important. As a direct result of this, an accuracy of 99.79 percent was gained for the dataset that was used for training, and an accuracy of 91.38 percent was acquired for the dataset that was used for testing. AptaNet reportedly achieved a satisfactory degree of performance on the benchmark dataset that we constructed by combining aptamers with proteins, as stated in [23]. According to the findings, AptaNet has the potential to help in the discovery of new aptamer–protein interaction pairs and the development of more effective insights into the link that exists between aptamers and proteins. Moreover, AptaNet has the potential to assist in the development of more effective insights into the link that exists between aptamers and proteins.

According to [24], utilizing computational approaches to produce accurate predictions of important proteins may help reduce the expense of doing research in wet labs. To construct protein-protein interaction (PPI) networks, the majority of the time, the currently available computational algorithms make use of several types of biological data. However, PPI networks and other types of biological data are not always of high quality for all proteins. Therefore, it is highly vital and valuable to develop methods that reliably predict essential proteins based just on their protein sequences. To increase the accuracy of determining which proteins are essential, [24] suggests using a machine learning ensemble model called EPGBDT. This model only considers protein sequences.

EPGBDT is different from other sequence-based predictors in two ways: By combining 49 GBDT base classifiers into a single ensemble model, I EP-GBDT can generate highly accurate and trustworthy predictions. (ii) EP-GBDT makes use of sampling to mitigate the impact of unbalanced data sets. EP-GBDT was evaluated by [24] using an independent test set, and it was compared to a sequence-based predictor known as Pheg, which is considered to be state-of-the-art. EPGBDT does well in all evaluation metrics and does better than Pheg. EP-GBDT is more accurate than the other 8 network-based centrality measures when compared further. All of the results show that EP-GBDT can be a useful tool for figuring out which human proteins are essential.

According to [25], the most challenging aspect of identifying and treating a neurological condition is locating the gene that is responsible for causing the disorder. In the field of biomedical research, it is very challenging to identify the specific genes that are responsible for the onset or progression of many disorders

that impact the nervous system, such as Parkinson's disease. Neurological illnesses are a significant component of genetics, and identifying them needs the use of techniques of machine learning that are still in the development stage. Since it is impractical to compare several sequences by hand, computational analysis is an essential technique for the study of protein sequences (genes). It makes it simple to find a gene in the sequence and organize the protein sequences that are connected into classes. There are a variety of tried-and-true diagnostic approaches that may be used to identify Parkinson's disease. However, there hasn't been nearly as much research conducted to compare different Machine Learning methods that make use of protein sequences to assess Parkinson's disease. In the article [25], a comparison is made between the many methods that may be used to categorize Parkinson's disease. These methods include examining the hydrophobicity of proteins as well as the amino acid composition of their proteins to extract characteristics. The rate of incorrect predictions is used to create a 2-level ensemble approach, which is then used to classify methods that have been combined. The efficacy of these approaches may be evaluated using metrics like precision, recall, F-score, and ROC curves. Under 5-fold cross-validation, experimental findings have demonstrated that the classifiers Random Forest, SVM, Neural Network (PCANNET), and Naive Bayes each performed the best based on their respective performance criteria. The suggested technique, on the other hand, beats Random Forest and SVM by 1.96 percentage points, NB by 1.1 percentage points, and PCANNET by 1.68 percentage points, respectively.

To predict potential RNA-aptamer candidates based on the known sequence of a target protein, the authors of this work [26] developed a model that they refer to as the Apta-MCTS. Recent research on nucleotide sequence classification has mostly concentrated on binary classification, but very little effort has been made to find acceptable aptamers. [26] devised a method for using machine learning to create candidate ribonucleic acid aptamers. This method is based on a classifier that can differentiate between API and MCTS. [26] ensured that our model extracted the appropriate characteristics from the input data by using the TPC and PseKNC encoders. With the aid of the API classifiers, which were based on the random forest model, the needed scores on the MCTS were established. [26] simulated how effectively their candidate aptamers and target proteins would bind to one other based on the molecular structures of both of them using ZDOCK. They were able to determine how effectively Apta-MCTS functioned as a result of this. The docking scores that were produced by Apta-MCTS were, on average, greater than those that were produced by known aptamers, and when compared to the results that were produced by other methods of creation, they were also higher than the results that were produced by known aptamers.

The models that were constructed by [26] have the potential of generating aptamer sequences for users to create that is of any length that may be requested by the user. [26] did some studies to find out how the length of aptamers influenced the various target proteins they were looking for. Those aptamers with 70–90 base pairs exhibited improved docking scores when compared to those aptamers that had been found before. All of these data demonstrate that their Apta-MCTS may generate aptamer sequences that are more suited for the studies at hand in contrast to other approaches that are presently being used in the field. These methods include:

In the paper [27], the authors provide the first computational method that can predict protein structures routinely and with atomic accuracy. This is possible even in the absence of a structure that is comparable to the protein in question. In the challenging 14th Critical Assessment of protein Structure Prediction (CASP14)15, the authors of their model, AlphaFold, which is based on neural networks, were able to demonstrate the validity of a whole new version of their model. This model displayed accuracy that was comparable with experimental structures in the majority of situations and greatly outperformed other techniques. Moreover, it significantly outperformed other methods. The most current version of AlphaFold makes use of an innovative method of machine learning as its foundation. This approach makes use of various sequence alignments to construct the deep learning algorithm with the help of physical and biological information on the construction of proteins.

The paper [28] talks about aptamers and reveals how their model is superior to others. Traditional drug development, as stated by [28], has centered on the antibody, the production of which requires a significant amount of time and effort. Several novel types of biomaterial, such as aptamers, which are short oligonucleotides with a single strand and a three-dimensional structure, have been produced as part of an effort to hasten the drug development process. Aptamers have a binding affinity that is comparable to that of antibodies, but they are less expensive and can be produced more quickly. An in vitro experimental technique known as systematic evolution of ligands by exponential enrichment, or SELEX, can be used to find aptamers that bind a certain target protein. The SELEX experiment must be carried out for its entirety

over several months. To cut down on the amount of time and money required for SELEX, various studies have been conducted to locate aptamers in silico; nonetheless, the majority of these studies concentrate on the interpretation of the findings of SELEX experiments. Some studies use machine learning to predict the interaction between aptamers and a target protein; however, these studies only feed the primary structure of the aptamers and proteins into their machine model, even though both aptamers and proteins exist in a three-dimensional space. Because of this, information is lost. [28] provide a new machine learning model that is based on a Transformer and that accepts as inputs aptamers and proteins in secondary structure. [28] validate their model by using benchmark datasets and comparing it to four different methodologies that are already in use. Their model performs better than others in this evaluation. In this publication, the authors state that they believe their model can increase the effectiveness of SELEX trials.

According to [20], the process of developing new pharmaceuticals is infamously difficult and expensive, with a low success rate. One of the most important activities that have to be done in the early stages of both the process of discovering new drugs and the process of repurposing existing drugs is the identification of drug-target interactions. A high binding affinity indicates that there is a significant interaction between the pharmaceutical and the target that it is intended to treat. In this regard, several different computational methods have been developed to predict the drug-target binding affinity, and it has been demonstrated that the input representation of these models is particularly effective in enhancing accuracy. In addition, several different computational methods have been developed to predict the drug-target binding affinity. Even while more recent models predict binding affinity with a better degree of precision than older models did, these models still need the three-dimensional structure of target proteins to be accurate. Even though there is a lot of interest in protein structure, there is a significant gap between the sequences of proteins that are already known and the structures of proteins that have been discovered via research. It is vital to locate an appropriate presentation for both the drug and the protein sequences to make an accurate prediction about the potential of the treatment to attach to its intended target. The fundamental purpose of this specific piece of study [20] is to assess the drug and protein sequence representation to improve the drug-target binding affinity prediction.

According to [29], aptamers are ligands that are formed of single-stranded nucleic acid, and they can attach to their targets with a very high degree of specificity and affinity. The great majority of the time, you'll be able to locate them by looking through several libraries for sequences that have excellent binding properties. On the other hand, these libraries can only access a tiny fraction of the whole sequence space that is conceptually conceivable. The use of machine learning makes it possible to intelligently navigate this area to discover aptamers that function exceptionally well. This opens up the opportunity. The authors of this research [29] present a strategy in which (PD) particle display is used to sort an aptamer library according to affinity, and then this information is used to train machine learning models to predict affinity in silico. Their method successfully predicted high-affinity DNA aptamers from experimental candidates at a rate that was 11 times greater than that of random perturbation. Additionally, it developed new high-affinity aptamers at a rate that was higher than what was observed when PD was employed by itself. The approach that they followed also made it simpler to construct truncated aptamers that were 70 percent shorter and had a higher binding affinity (1.5 nM) than the best experimental candidate. By combining machine learning with physical methods, as shown in this study, it is feasible to accelerate the creation of improved diagnostic and therapeutic medications.

For a complete comprehension of a variety of physiological processes, such as signal cascades, DNA transcription, metabolic cycles, and cellular repair, it is vital to have an understanding of protein-protein interactions, which are also referred to by their acronym, PPIs. Over the last decade, a great deal of research has gone into the development of high-throughput methods for locating PPIs. Despite this, these techniques call for a significant investment of time and labor, and they virtually always provide a significant percentage of incorrect negative results. As a consequence of this, there is a substantial need for the development of cutting-edge computational algorithms that are capable of acting as additional tools for PPI prediction. [30] presents an innovative sequence-based approach to the problem of predicting PPIs. The Discrete Hilbert transform (DHT) and the Rotation Forest are also included in this model. The whole of this procedure may be broken down into three separate stages, and they are as follows: In the beginning, the Position-Specific Scoring Matrices (PSSM) approach was used to transform the amino acid sequence into a PSSM matrix. PSSM stands for position-specific scoring matrices. The history of proteins may be stored in this matrix, which can hold a tremendous quantity of data. After that stage was finished being worked on, the next thing that was done was to generate a DHT description in 400 dimensions for every

possible pair of proteins. In the conclusion, the RoF classifier was used to establish the most likely PPI class by making use of the feature descriptors that were supplied. During the study, we were able to use the model that had been suggested to obtain remarkable accuracies of 91.93, 96.35, and 94.24 percent, respectively, for the PPIs datasets for yeast, humans, and Oryza sativa, respectively. These outcomes could only be accomplished by using the data sets that were gathered. In addition, [30] has conducted a large number of tests using PPI datasets that span many species. They concluded that the predictive power of our approach is likewise of an exceptionally high standard. [30] compare the results of RoF with those of four other sophisticated classifiers, namely the Support Vector Machine (SVM), Random Forest (RF), K-Nearest Neighbor (KNN), and AdaBoost. In addition to that [30], Existing works are already of a higher quality when compared to those of other authors. These exhaustive experimental findings provide additional validation for the excellence of the approach that has been suggested, as well as its practicability. [Citation needed] They anticipate that it will be useful to them as a supplemental instrument for proteomics analysis in their ongoing and upcoming studies.

The treatment of some forms of cancer has undergone a sea shift in recent years as a result of the introduction of immune checkpoint-targeted immunotherapy. It could be easier to conclude if the condition of immune checkpoint expression in certain cancers has been established. In this article [31], the design and development of a molecular probe that detects human PD-L1 with high specificity are discussed. The probe is based on a single-stranded aptamer, and it was designed to target the protein. Following the selection of target-engaging aptamers from a pool of random DNA through an iterative enrichment procedure, the binding is characterized by biochemical methods. Specificity and dosage dependency were proven in vitro in a cell culture setting using human kidney tumor cells (786-0), human melanoma cells (WM115 and WM266.4), and human glioblastoma LN18 cancer cells. [31] reveals that the probe divulges good potential in imaging, which proves that the probe is beneficial in vivo by using two mouse tumor models. [31] reveals that the probe reveals excellent potential in imaging. [31] theorizes that possible improvements to the probe soon might make it possible to do universal imaging of many kinds of tumors based on the PD-L1 status of the tumors, which could be useful in the process of detecting cancer. The most current research on predicting aptamers is summarised in Table 1, which may be found here.

Table 1. Comparative study of literature for the prediction of new aptamers using ML techniques.

Authors	Year	Objective	Dataset Description	Models	Target	Results
[19]	2020	Predicting of the sequence RNA-binding proteins	31 RBP datasets	RBPCNN	RNA	The average area under the receiver operator curve was improved by 2.67 percent and the mean average precision was improved by 8.03 percent.
[20]	2021	Predict RNA sequences And Produce new sequences	Protein Data Bank (PDB)	CD13	To Predict RNA Sequences	Accuracy = 92.72%
[21]	2022	Target and Produced the new Aptamers against SGIV Infrctionn	The data that support the findings of this study are available from the	AHTS	Feature selection	-

			corresponding author upon reasonable request.			
[23]	2021	Predict the aptamer–protein interaction pairs by integrating features derived from both aptamers and the target proteins	Aptagen & Aptamer Base	AptaNet	To Target Protein	99.79% accuracy was achieved for the training dataset, and 91.38% accuracy was obtained for the testing dataset
[24]	2022	To predict the essential proteins by using only protein sequences	Database of Essential Genes (DEG) https://tubic.org/deg/public/index.php	EP-GBDT	To Target the essential protein	-
[26]	2021	To determine potential RNA-aptamer candidates for a target protein	Aptamers base & Protein Data Bank (PDB)	Apta-MCTS	To Target RNA Aptamers	-
[27]	2021	Predict protein structures				

3. Materials and Methods

This section contains the experimental process which was conducted to generate the strong aptamers candidates from the RNA-protein complexes.

3.1 Datasets Description

We collected the P-RNA [32-34 complexes dataset from [19] having a resolution of 5.0 Å which has been solved by using X-ray crystallography (XRC). XRC is used in experimental science to determine the atomic and molecular composition of the crystal. The dataset containing the size of RNA sequences having less than 10 or greater than 120 nucleotides was removed and not considered for this study. Therefore, the dataset only contains a total of 696 P-RNA complexes, which were applied to observe the computational analysis of amino acids with nucleotides. In addition, the P-RNA sequences consist of 22 categories based on the protein data bank (PDB) [35-38]. The proposed model was also applied to the publically available benchmark dataset [13], which was obtained from the aptamer database [20]. This dataset consists of a total of 580 DNA or P-RNA pairs including (145 positive and 435 negative sequences). We select 100 RNA aptamers protein instances of positive and negative sequences for evaluating our proposed model, RF, and SVM [39-41].

3.2 Proposed MLP

Our methodology for finding the RNA aptamers consists of two sections: training of a neural network (MLP) was performed by extracting the RNA key features such as mC, dC, PseTNC & PseAAC, and feeding it to the MLP, and then predicting the aptamers candidates to target the protein as illustrated in Figure 1.

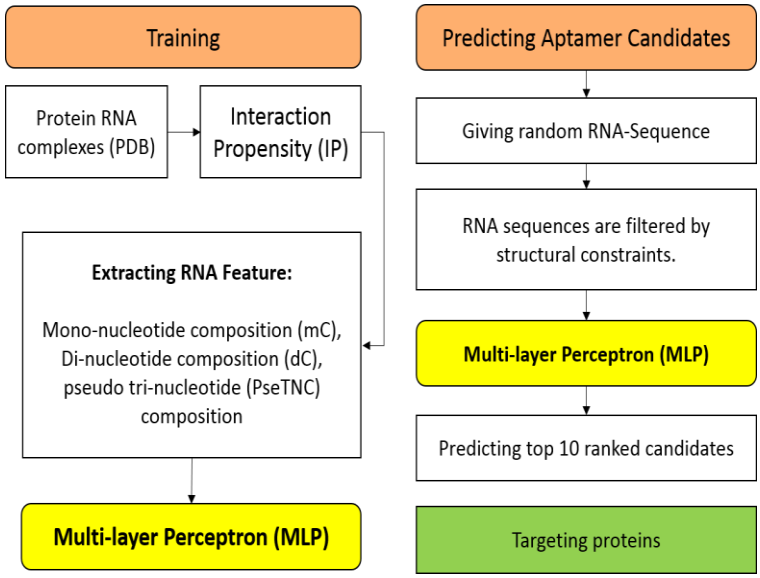

Figure 1. Framework for predicting potential RNA Aptamers

The purpose of training the MLP [42] was to calculate the probability measure of RNA sequences using their key features. For predicting the aptamer candidates (see Figure 1), we produced the random RNA sequences of length 25-mer to find their secondary structure. The reason behind selecting the 25-mer RNA sequence is that; the SELEX process selects the aptamers of size 30-mer to 60-mer from the oligonucleotide libraries [43-46]. Though, aptamers are normally smaller than the size of 30-mer [21-22]. Therefore, the main focus of this study was to produce 25-mer RNA aptamers to target the protein. In the end, we sorted the aptamer candidates [47-49] in descending order, and then top-ranked candidates have been docked to target the protein. The comprehensive description of our work is summarized in Algorithm 1.

Algorithm 1:
Predicting potential RNA aptamers candidates to target protein using neural network

Input: protein target ***pt***, protein-RNA complexes (***P-RNA***), interaction propensity (***IP***).
Output: Predicting the aptamers candidates to the target protein.
S ← ∅ **{S: consist of P-RNA complexes from PDB}**
IP ← extracting RNA feature mC, dC & PseTNC of an amino acid from ***P-RNA*** complexes.
Train the proposed **MLP** using ***IP***
Apply **RRC** ← Random RNA complexes (not greater than **25-mer**)
foreach candidate c ∈ S **do**
    f ← feature vector of c with positive and negative ***pt*** instances
    c. Probability (positive votes of MLP)
    According to their probability, sort c in descending order.
. **End**

Most of the research observes that the interaction of nucleotide triplets with amino acids is the most important aspect to consider when attempting to anticipate the P-RNA interaction [23-25]. [50-51] In addition, we measured the mC, dC, and PseTNC [26] for every RNA sequence by using a database that included 696 P-RNA complexes. The value of PseTNC was determined by the use of three physiochemical parameters, namely hydrophobicity (H), hydrophilicity (HP), and side-chain mass (SCM), respectively [27], [28], and [29]. The initial five computed value of PseTNC has been discussed in Table 2. Then, we clustered the 20 amino acids {A, R, N, D, C, Q, E, G, H, I, L, K, M, F, P, S, T, W, Y, V} into 7 respective groups based on dipole of the chain. These groups were {M,P,S,T}, {N,D,C,Q}, {A,L,R,K}, {F,W,Y}, {E,I}, {G,V} and {H}. The purpose behind clustering the amino acid into seven groups was to decrease the length of a feature vector to denote the protein sequence. In addition, the clustering of amino acids into seven sets was also applied successfully in multiple research studies [30-32].

Table 2. PseTNC values

No	TNC	H	HP	SCM
1	AAT	-0.78	0.2	58
2	ATA	1.38	-1.8	57
3	CAA	-0.75	0.2	72
4	CCA	0.12	0	42
5	ACG	-0.04	-0.5	40

3.3 Positive (+) and negative (-) instances for training

As we discussed earlier, the MLP model was trained on the feature vector obtained from the P-RNA complexes. The MLP model consists of 16 layers, and the feature for extracting RNA sequences was set to the square root of the feature elements [52-59]. We compared the performance of our model with two well-renowned machine classifiers i.e. RF and SVM. However, grid search was applied to determine the parameters for both of the machine learning classifiers (RF and SVM). The objective of our model is to generate the potential 25-mer RNA aptamers, but the RNA sequences provided at the time of training was of different length [60-67]. Figure 2 represents the sliding window of 25 nucleotides for positive and negative instances separately. In addition, the positive (+) symbol shows the protein-binding nucleotide, while the negative (-) sign denotes the non-binding nucleotide. The window is considered positive if the middle of the window binds the protein nucleotide with (+) instance, and the window containing non-binding nucleotide is considered negative.

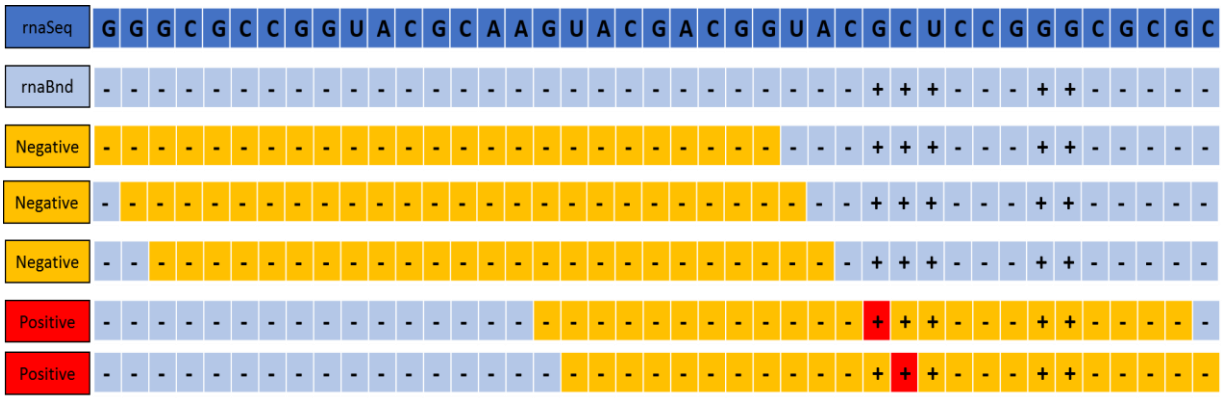

Figure 2. Positive and negative windows of 25 nucleotides in RNA sequences

We also removed the feature vectors that were neither (+) nor (-) from the training phase because it may produce severely unbalanced instances for training, unless and until the initial and final windows were supposed as (+) if they consist of a protein-binding nucleotide in any location of the sliding window [68-72]. This is due to the limited number of (+) instances than (-) in the training database. Therefore, the ratio of (+) and (-) instances at the time of training are about 1:3 [73-75].

3.4 Filtering RNA sequences by structural constraint

Initially, we produced the random RNA sequences to find the 25-mer RNA aptamers and then applied RNAfold to predict their secondary structures [33]. The secondary structure of the RNA sequences must contain free energy lower than -5.7kcal/mol and their pool should not be greater than 150. The process of secondary structure was used to develop the pool of aptamer candidates. The limitation of the free energy was selected from the research study presented by [10]. All of these generated pools of aptamer candidates were applied to the MLP model, and then select the top 10 ranked potential aptamers based on their probability and free energy. In addition, HDOCK [34] was also used to evaluate the performance of docking the potential aptamers to target the protein [76].

3.5 Cross-validation and leave-one-out validation

The performance of the model was evaluated by using 15-cross validation (CV) [77] and leave-one-out (LOO) validation. Cross-validation (CV) [78] is very significant when the amount of data is scarce splitting the dataset into training and testing sections. Although the leave-one-out validation returns the identical output for an individual dataset [35-36]. The performance of the MLP, RF, and SVM was evaluated by 6 metrics such as SN, SP, ACC, MCC, PPV, and NPV which were calculated by the following equations (1-6).

$$SN = \frac{TP}{TP+FN} \quad (1)$$

$$SP = \frac{TN}{TN+FP} \quad (2)$$

$$ACC = \frac{TP+TN}{TP+TN+FP+FN} \quad (3)$$

$$MCC = \frac{(TP*TN)-(FP*FN)}{(\sqrt{TP+FP)*(TP+FN)*(TN+FP)*(TN+FN)})} \quad (4)$$

$$PPV = \frac{TP}{TP+FP} \quad (5)$$

$$NPV = \frac{TN}{TN+FN} \quad (6)$$

4. Results

This section contains a detailed description of the experiment performed to predict the potential aptamers by using the P-RNA complexes database. The performance of the MLP, RF, and SVM models has been also evaluated in this section. Moreover, the result obtained by applying these approaches to the independent benchmark dataset created by Li et al. [13] was also part of this section.

4.1 Experimental setup

In this phase, we extract the RNA features before feeding the model was performed on Microsoft Visual Studio in C# programming language [35]. The MLP RF & SVM models were deployed with the help of the Keras framework and sci-kit learn repository respectively in Python [37]. The research experiment was executed on a Window based operating system with 11GB GPU NVIDIA GeForce GTX and 32GB RAM.

4.2 Potential Aptamers with P-RNA complexes

The RNA sequences that were obtained from the pool following the application of structural restrictions were gathered. To target the possible aptamer protein, this pair of sequences was stored in the feature vector for the MLP model. The MLP model had sixteen layers, and the input feature vector was the feature vector of the P-RNA complexes. After that, the probability of the positive vector was computed, and then we chose the RNA sequence that had a greater probability and had the free energy values that were the lowest. HDOCK was used to carry out the docking of the top 10 aptamers that were designed to target the protein. This allowed for the structure of the expected aptamers to be seen. As can be seen in Figure 3, the border structure of the anticipated aptamers was generated with the help of the RNAComposer [38]. It has been observed that the blueprints of the top 10 anticipated candidates for RNA aptamers target proteins with the same place on their structures as the blueprints of the RNA aptamers that are delivered.

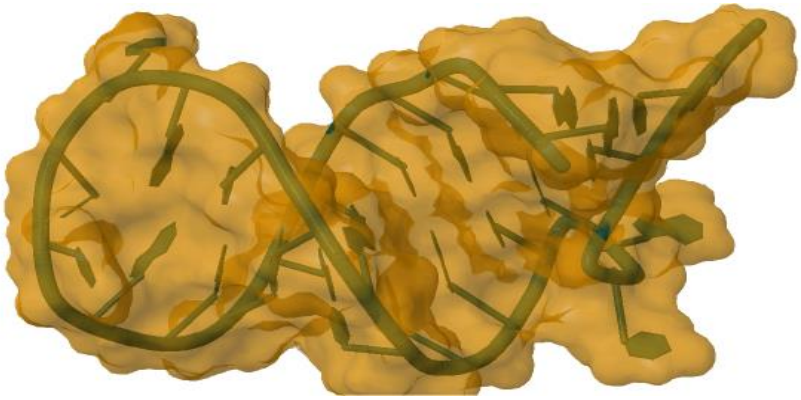

Figure 3. Boundary structure of the aptamers to a target protein (PDB ID: 3DD2)

We also compared the 25-mer RNA aptamers produced by our MLP method with the longer RNA aptamers having more than 25 nucleotides. Figure 4 represents the comparison of our generated aptamers with the large RNA aptamers. Even with the difference between the length of the predicted and actual aptamers, both of the RNA aptamers represent very similar binding tertiary

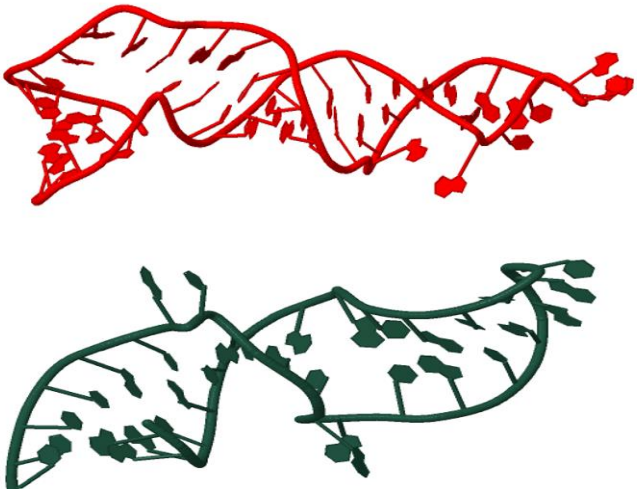

Figure 4. Comparison of 25-mer RNA aptamers (red) with the large size of RNA aptamers (33-mer nucleotides).

In addition, the results that were produced by the MLP, RF, and SVM after using two different validation methods are shown in Table 3, as seen above. The findings of the CV validation were much more favorable than those of the LOO approach, even though both validations (CV and LOO) demonstrate high performance in terms of six different metrics. In terms of all assessment measures, the MLP model demonstrates remarkability in its performance. Additionally, in comparison to the MLP and RF models, the PPV value predicted by the SVM model is predicted to be lower. It seems that the SVM model has a greater number of false positive values than the MLP model and the RF model.

Table 3. The output of 15-fold CV and LOOCV of the MLP, RF, and SVM

Validation	PPV (%)	NPV (%)	SN (%)	SP (%)	ACC (%)	MCC (%)
15-fold (MLP)	97.58	97.47	96.56	97.52	98.44	91.23
15-fold (RF)	96.10	95.62	94.22	98.35	96.33	89.10
15-fold (SVM)	90.12	96.32	93.54	92.56	90.78	88.65
LOO (MLP)	96.50	93.93	96.12	98.25	98.10	93.54
LOO (RF)	95.01	93.26	94.42	96.56	97.79	92.22
LOO (SVM)	90.02	92.99	94.98	95.19	96.12	89.06

We also used our MLP model to the benchmark dataset that was created by Li et al. [13] to conduct an independent performance evaluation of it. In their study, Zhang et al. [14] also made use of the same dataset [13]. Because of this, we additionally evaluate the findings of our model in light of the findings of these two previous pieces of research [13][14]. The benchmark dataset is made up of negative cases that were produced by using a random mix of aptamers to target the protein found in the dataset's positive instances [80]. So, we collected both positive and negative instances from it and applied them to our model.

As can be seen in Table 4, our MLP model managed to attain an SN of 75.8%, SP of 68.23%, ACC of 77.5%, PPV of 69.9%, NPV of 74.12%, and MCC of 39.9%. In terms of SN (48.3%), ACC (77.4%), and PPV (55.6%), the result that was provided by [13] demonstrates a lower level of performance. The research [14] showed better findings [13] and increased the values of SN and SP to 73.8% and 71.3% respectively from their previous levels. However, the PPV result that was published by [14] (46.1%), was lower than the one that was reported by Li et al [13]. In comparison to the other two trials, our model MLP was able to generate much better results in terms of SN, ACC, PPV, and MCC. Our MLP model had an SP that was 72.23%, which was a lower value than the one found by Li et al. [13], but a higher value than the one found by Zhang et al. [14]. Furthermore, the PPV of our model is ten times greater than that of Li's technique.

Table 4. Independent testing of the MLP with benchmark dataset.

Ref	Method	SN (%)	SP (%)	ACC (%)	PPV (%)	NPV (%)	MCC (%)
[13]	RF	48.3	87.1	77.4	55.6	83.5	37.2
[14]	Ensemble	73.8	71.3	71.9	46.1	89.1	39.8
Proposed Method	MLP	75.2	72.23	77.5	69.9	74.12	39.9

5. Discussion

This part offers an in-depth study of the output that was created by the suggested MLP model based on six performance assessment parameters. These parameters are as follows: SN, SP, ACC, PPV, and NPV, as well as MCC. This research study is comprised of five different processes: the acquisition of data from the PDB, the extraction of RNA characteristics, the training model, the prediction of probable aptamers, and the analysis of the results. One dataset was retrieved from the PDB, while the other was a benchmark dataset taken from [13][70]. During the training process, many RNA characteristics were retrieved for the model. The accuracy of our suggested model in predicting possible aptamers was 98.44% based on 15-fold cross-validation, and it was 98.10% based on LOO, which suggests that MLP is more effective than those of the other two models, RF and SVM (see Table 3). As can be seen in Table 4, the performance of the MLP on the other benchmark dataset was equally outstanding. It achieved an accuracy of 77.5% and a sensitivity of 75.2%, both of which are higher than those of the other two experiments. According to the results of the performance study of RF, the MLP model is superior to various other models [65-72]. We performed a 15-fold LOO to further test the outcomes of the proposed model, and the result validates the relevance of MLP, as can be shown in Table 3. The findings also indicate that our model is useful for making accurate predictions about the new aptamers that will be used to target the protein.

6. Conclusion

The search for aptamers has made extensive use of several different computational methods. The majority of the research cannot be used in the process of discovering new possible aptamers to target the protein since the primary purpose of these studies was to determine whether or not a certain pair of RNA sequences and protein interact with one another. As a consequence of this, we devised an innovative computational technique, which we put to use to construct the prospective RNA aptamers that target the protein by extracting the various characteristics of the interacting RNA. We construct and train an MLP model by using several different characteristics of P-RNA sequences. Even though it is still in its early stages, the MLP model has shown promising results in the cross-validation approaches as well as the independent testing on the benchmark dataset. We believe that our approach will be beneficial in lowering the amount of time and money spent on in vitro testing, as well as useful in reducing the main size of the nucleic acid sequence pool.

Funding: This research received no external funding.
Data Availability Statement: The authors declare that all data supporting the findings of this study are available within the article.
Acknowledgment: We are thankful to the Research & Development Department of the National College of Business Administration & Economics Lahore, Multan Sub-Campus for providing us with the lab and other resources to complete this project effectively.
Conflicts of Interest: The authors declare no conflict of interest.

References

1. Song, K. M., Lee, S., & Ban, C. (2012). Aptamers and their biological applications. *Sensors*, *12*(1), 612-631.
2. Tuerk, C., & Gold, L. (1990). Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *science*, *249*(4968), 505-510.
3. DONG, H. Y., WANG, J., ZHANG, T., MA, J., HAN, L. Y., TU, X. H., & JIA, L. (2016). Aptamers and their biological applications in bio-medicine and analytical diagnostics. *Chinese Journal of Pharmaceutical Analysis*, *36*(3), 369-376.
4. Ellington, A. D., & Szostak, J. W. (1990). In vitro selection of RNA molecules that bind specific ligands. *nature*, *346*(6287), 818-822.
5. Li, H., Jin, H., Wan, W., Wu, C., & Wei, L. (2018). Cancer nanomedicine: mechanisms, obstacles and strategies. *Nanomedicine*, *13*(13), 1639-1656.
6. He, J., Wang, J., Zhang, N., Shen, L., Wang, L., Xiao, X., ... & Shangguan, D. (2019). In vitro selection of DNA aptamers recognizing drug-resistant ovarian cancer by cell-SELEX. *Talanta*, *194*, 437-445.
7. Vidic, M., Smuc, T., Janez, N., Blank, M., Accetto, T., Mavri, J., ... & Lah, T. T. (2018). selection approach to develop DNA aptamers for a stem-like cell subpopulation of non-small lung cancer adenocarcinoma cell line A549. *Radiology and oncology*, *52*(2), 152-159.
8. Su, Y., Xu, H., Chen, Y., Qi, J., Zhou, X., Ge, R., & Lin, Z. (2018). Real-time and label-free detection of bisphenol A by an ssDNA aptamer sensor combined with dual polarization interferometry. *New Journal of Chemistry*, *42*(4), 2850-2856.
9. Yang, B., Bao, W., Huang, D. S., & Chen, Y. (2018). Inference of large-scale time-delayed gene regulatory network with parallel mapreduce cloud platform. *Scientific Reports*, *8*(1), 1-11.
10. Chushak, Y., & Stone, M. O. (2009). In silico selection of RNA aptamers. *Nucleic acids research*, *37*(12), e87-e87.
11. Osborne, S. E., & Ellington, A. D. (1997). Nucleic acid selection and the challenge of combinatorial chemistry. *Chemical reviews*, *97*(2), 349-370.
12. Bao, W., Yang, B., Li, Z., & Zhou, Y. (2018). LAIPT: lysine acetylation site identification with polynomial tree. *International Journal of Molecular Sciences*, *20*(1), 113.
13. Li, B. Q., Zhang, Y. C., Huang, G. H., Cui, W. R., Zhang, N., & Cai, Y. D. (2014). Prediction of aptamer-target interacting pairs with pseudo-amino acid composition. *PLoS One*, *9*(1), e86729.
14. Zhang, L., Zhang, C., Gao, R., Yang, R., & Song, Q. (2016). Prediction of aptamer-protein interacting pairs using an ensemble classifier in combination with various protein sequence attributes. *BMC bioinformatics*, *17*(1), 1-13.
15. Hu, W. P., Kumar, J. V., Huang, C. J., & Chen, W. Y. (2015). Computational selection of RNA aptamer against angiopoietin-2 and experimental evaluation. *BioMed research international*, *2015*.
16. Shcherbinin, D. S., Gnedenko, O. V., Khmeleva, S. A., Usanov, S. A., Gilep, A. A., Yantsevich, A. V., ... & Archakov, A. I. (2015). Computer-aided design of aptamers for cytochrome p450. *Journal of structural biology*, *191*(2), 112-119.
17. Ahirwar, R., Nahar, S., Aggarwal, S., Ramachandran, S., Maiti, S., & Nahar, P. (2016). In silico selection of an aptamer to estrogen receptor alpha using computational docking employing estrogen response elements as aptamer-alike molecules. *Scientific reports*, *6*(1), 1-11.
18. Rabal, O., Pastor, F., Villanueva, H., Soldevilla, M. M., Hervas-Stubbs, S., & Oyarzabal, J. (2016). In silico aptamer docking studies: from a retrospective validation to a prospective case study'tim3 aptamers binding. *Molecular Therapy-Nucleic Acids*, *5*, e376.
19. Tayara, H., & Chong, K. T. (2020). Improved predicting of the sequence specificities of RNA binding proteins by deep learning. IEEE/ACM transactions on computational biology and bioinformatics, 18(6), 2526-2534.
20. Torkamanian-Afshar, M., Nematzadeh, S., Tabarzad, M., Najafi, A., Lanjanian, H., & Masoudi-Nejad, A. (2021). In silico design of novel aptamers utilizing a hybrid method of machine learning and genetic algorithm. Molecular diversity, 25(3), 1395-1407.
21. Wei, H., Liu, M., Ke, K., Xiao, S., Huang, L., He, Q., ... & Yu, Q. (2022). Study on aptamer based high throughput approach identifies natural ingredients against RGNNV. Journal of Fish Diseases.
22. Shaath, H., Vishnubalaji, R., Elango, R., Kardousha, A., Islam, Z., Qureshi, R., ... & Alajez, N. M. (2022, May). Long Non-Coding RNA and RNA-Binding Protein Interactions in Cancer: Experimental and Machine Learning Approaches. In Seminars in Cancer Biology. Academic Press.
23. Emami, N., & Ferdousi, R. (2021). AptaNet as a deep learning approach for aptamer–protein interaction prediction. Scientific Reports, 11(1), 1-19.
24. Wang, T., Zhang, H., Wu, Y., Jiang, W., Chen, X., Zeng, M., ... & Yang, Z. (2022). Target discrimination, concentration prediction, and status judgment of electronic nose system based on large-scale measurement and multi-task deep learning. Sensors and Actuators B: Chemical, 351, 130915.

25. Arora, P., Mishra, A., & Malhi, A. (2022). Machine learning Ensemble for the Parkinson's disease using protein sequences. Multimedia Tools and Applications, 1-28.
26. Lee, G., Jang, G. H., Kang, H. Y., & Song, G. (2021). Predicting aptamer sequences that interact with target proteins using an aptamer-protein interaction classifier and a Monte Carlo tree search approach. PloS one, 16(6), e0253760.
27. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., ... & Hassabis, D. (2021). Applying and improving AlphaFold at CASP14. Proteins: Structure, Function, and Bioinformatics, 89(12), 1711-1721.
28. Shin, I., & Song, G. (2022, January). Aptamer-Protein Interaction Prediction using Transformer. In 2022 IEEE International Conference on Big Data and Smart Computing (BigComp) (pp. 368-370). IEEE.
29. Bashir, A., Yang, Q., Wang, J., Hoyer, S., Chou, W., McLean, C., ... & Ferguson, B. S. (2021). Machine learning guided aptamer refinement and discovery. Nature communications, 12(1), 1-11.
30. Pan, J., Wang, S., Yu, C., Li, L., You, Z., & Sun, Y. (2022). A Novel Ensemble Learning-Based Computational Method to Predict Protein-Protein Interactions from Protein Primary Sequences. Biology, 11(5), 775.
31. Malicki, S., Pucelik, B., Żyła, E., Benedyk-Machaczka, M., Gałan, W., Golda, A., ... & Dubin, G. (2022). Imaging of Clear Cell Renal Carcinoma with Immune Checkpoint Targeting Aptamer-Based Probe. Pharmaceuticals, 15(6), 697.
32. Lee, W., & Han, K. (2019). Constructive prediction of potential RNA aptamers for a protein target. IEEE/ACM transactions on computational biology and bioinformatics, 17(5), 1476-1482.
33. Cruz-Toledo, J., McKeague, M., Zhang, X., Giamberardino, A., McConnell, E., Francis, T., ... & Dumontier, M. (2012). Aptamer base: a collaborative knowledge base to describe aptamers and SELEX experiments. *Database*, *2012*.
34. Ruckman, J., Green, L. S., Beeson, J., Waugh, S., Gillette, W. L., Henninger, D. D., ... & Janjic, N. (1998). 2′-Fluoropyrimidine RNA-based aptamers to the 165-amino acid form of vascular endothelial growth factor (VEGF165): Inhibition of receptor binding and VEGF-induced vascular permeability through interactions requiring the exon 7-encoded domain. *Journal of Biological Chemistry*, *273*(32), 20556-20567.
35. Tasset, D. M., Kubik, M. F., & Steiner, W. (1997). Oligonucleotide inhibitors of human thrombin that bind distinct epitopes. *Journal of molecular biology*, *272*(5), 688-698.
36. Choi, S., & Han, K. (2011, December). Prediction of RNA-binding amino acids from protein and RNA sequences. In *Bmc Bioinformatics* (Vol. 12, No. 13, pp. 1-12). BioMed Central.
37. Choi, S., & Han, K. (2013). Predicting protein-binding RNA nucleotides using the feature-based removal of data redundancy and the interaction propensity of nucleotide triplets. *Computers in Biology and Medicine*, *43*(11), 1687-1697.
38. Tuvshinjargal, N., Lee, W., Park, B., & Han, K. (2015). Predicting protein-binding RNA nucleotides with consideration of binding partners. *Computer Methods and Programs in Biomedicine*, *120*(1), 3-15.
39. Chen, W., Feng, P. M., Deng, E. Z., Lin, H., & Chou, K. C. (2014). iTIS-PseTNC: a sequence-based predictor for identifying translation initiation site in human genes using pseudo trinucleotide composition. *Analytical biochemistry*, *462*, 76-83.
40. Tanford, C. (1962). Contribution of hydrophobic interactions to the stability of the globular conformation of proteins. *Journal of the American Chemical Society*, *84*(22), 4240-4247.
41. Hopp, T. P., & Woods, K. R. (1981). Prediction of protein antigenic determinants from amino acid sequences. *Proceedings of the National Academy of Sciences*, *78*(6), 3824-3828.
42. Kemmei, T., Kodama, S., Yamamoto, A., Inoue, Y., & Hayakawa, K. (2015). Reversed phase liquid chromatographic determination of organic acids using on-line complexation with copper (II) ion. *Analytica Chimica Acta*, *886*, 194-199.
43. You, Z. H., Chan, K. C., & Hu, P. (2015). Predicting protein-protein interactions from primary protein sequences using a novel multi-scale local feature representation scheme and the random forest. *PloS one*, *10*(5), e0125811.
44. Shen, J., Zhang, J., Luo, X., Zhu, W., Yu, K., Chen, K., ... & Jiang, H. (2007). Predicting protein–protein interactions based only on sequences information. *Proceedings of the National Academy of Sciences*, *104*(11), 4337-4341.
45. Zhou, X., Park, B., Choi, D., & Han, K. (2018). A generalized approach to predicting protein-protein interactions between virus and host. *BMC genomics*, *19*(6), 69-77.
46. Lorenz, R., Bernhart, S. H., Höner zu Siederdissen, C., Tafer, H., Flamm, C., Stadler, P. F., & Hofacker, I. L. (2011). ViennaRNA Package 2.0. *Algorithms for molecular biology*, *6*(1), 1-14.
47. Yan, Y., Zhang, D., Zhou, P., Li, B., & Huang, S. Y. (2017). HDOCK: a web server for protein–protein and protein–DNA/RNA docking based on a hybrid strategy. *Nucleic acids research*, *45*(W1), W365-W373.
48. Choi, D., Park, B., Chae, H., Lee, W., & Han, K. (2017). Predicting protein-binding regions in RNA using nucleotide profiles and compositions. *BMC systems biology*, *11*(2), 1-12.
49. Kim, B., Alguwaizani, S., Zhou, X., Huang, D. S., Park, B., & Han, K. (2017). An improved method for predicting interactions between virus and human proteins. *Journal of bioinformatics and computational biology*, *15*(01), 1650024.
50. Malik, H., Farooq, M. S., Khelifi, A., Abid, A., Qureshi, J. N., & Hussain, M. (2020). A comparison of transfer learning performance versus health experts in disease diagnosis from medical imaging. *IEEE Access*, *8*, 139367-139386.
51. Popenda, M., Szachniuk, M., Antczak, M., Purzycka, K. J., Lukasiak, P., Bartol, N., ... & Adamiak, R. W. (2012). Automated 3D structure composition for large RNAs. *Nucleic acids research*, *40*(14), e112-e112.
52. Hussain, Z., Imran, M., & Nosheen, A. (2022). Forensics to Government Agencies Data using Hyper Ledger Fabric (HLF). *Journal of Computing & Biomedical Informatics*, *3*(01), 208-229.
53. Iftikhar, F., Amin, H. M. M., & Abbas, G. (2022). A Feature Fusion Based Hybrid Approach for Breast Cancer Classification. Journal of Computing & Biomedical Informatics, 3(01), 243-256.
54. Tariq, H., & Iftikhar, A. (2022). Towards Developing Secure Transmission of Electronic Medical Records using Mobile Devices. Journal of Computing & Biomedical Informatics, 3(01), 257-266.

55. Faheem, M. R., Iftikhar, A., & Hussain, N. (2022). Automated Diagnosing of Eye Disease in Real Time. Journal of Computing & Biomedical Informatics, 3(01), 282-288.
56. Saima Anwar, Sana Akhtar, & Gra Badshah. (2022). Evaluation of Active Queue Management Methods based on Integrated AHP, TOPSIS and Fuzzy TOPSIS. Journal of Computing & Biomedical Informatics, 3(01), 267–281.
57. Javed, R., Khan, A. S., & Khan, A. H. (2022). Deep Learning Techniques for Diagnosis of Lungs Cancer. Journal of Computing & Biomedical Informatics, 3(01), 230-242.
58. Bashirpour Bonab, A., Fedele, M., Formisano, V., & Rudko, I. (2022). Quantum Technologies for Smart Cities: A Comprehensive Review and Analysis. Available at SSRN 4231189.
59. Farooqi, M. A., Ashraf, M. A., & Shaukat, M. U. (2021). Google Page Rank Site Structure Strategies for Marketing Web Pages. Journal of Computing & Biomedical Informatics, 2(02), 140-157.
60. Akram, A., Jiadong, R., Rizwan, T., Irshad, M., Noman, S. M., Arshad, J., & Badar, S. U. (2021). A Pilot Study on Survivability of Networking Based on the Mobile Communication Agents. International Journal of Network Security, 23(2), 220-228.
61. Aziz, O., Siraj, M. A., & Rehman, A. (2021). Privacy challenges in cyber security against cybercrime in digital forensic. A systematic literature review in Pakistan. Journal of Computing & Biomedical Informatics, 2(02), 158-164.
62. Fatima, S., Aziz, O., & Ahmad, M. U. (2021). Predictive Analysis about Traffic, Vehicles, and Road Congested Area: A Systematic Survey. *Journal of Computing & Biomedical Informatics*, *2*(02), 165-186.
63. Siddique, A., Zaidi, A. R., & Abbas, G. (2021). Analysis of 5th Generation Mobile Networks Architecture with Essential Wireless Access Technologies. *Journal of Computing & Biomedical Informatics*, *2*(02), 187-193.
64. Waqas, M., Imran, M., & Zaidi, A. R. (2021). A Novel Algorithm for Moving Target Detection. *Journal of Computing & Biomedical Informatics*, *2*(02), 194-207.
65. Hussain, A., Malik, H., & Chaudhry, M. U. (2021). Supervised Learning Based Classification of Cardiovascular Diseases. Journal: Proceedings of Engineering and Technology Innovation, 24-34.
66. Siraj, M. A., Rehman, A., Aziz, O., & Khan, M. F. (2021). Systematic Literature Review: Smart Drone for Early Smoke Detection in Forest Using IOT. *Journal of Computing & Biomedical Informatics*, *2*(01), 80-88.
67. Zaman, R., Bashir, R., & Zaidi, A. R. (2021). Image Classification and Text Extraction using Convolutional Neural Network. *Journal of Computing & Biomedical Informatics*, *2*(01), 89-95.
68. Abbas, Q., Imran, M., & Sajid, M. (2021). Integration of Healthcare Services of Specialized Healthcare & Medical Education Department, Government of Punjab in Cloud-based System. Journal of Computing & Biomedical Informatics, 2(01), 96-110.
69. Mahmood, R., Imran, M., & Hussain, S. K. (2021). Assessment of Network & Processor Virtualization in Cloud Computing. Journal of Computing & Biomedical Informatics, 2(01), 111-127.
70. Komal, A., & Malik, H. (2022). Transfer Learning Method with Deep Residual Network for COVID-19 Diagnosis Using Chest Radiographs Images. In Lecture Notes in Networks and Systems (pp. 145–159). Springer Nature Singapore. https://doi.org/10.1007/978-981-16-7618-5_13
71. Saeed, H., Malik, H., Bashir, U., Ahmad, A., Riaz, S., Ilyas, M., Bukhari, W. A., & Khan, M. I. A. (2022). Blockchain technology in healthcare: A systematic review. In P. Vijayakumar (Ed.), PLOS ONE (Vol. 17, Issue 4, p. e0266462). Public Library of Science (PLoS). https://doi.org/10.1371/journal.pone.0266462
72. Malik, H., Bashir, U., & Ahmad, A. (2022). Multi-classification neural network model for detection of abnormal heartbeat audio signals. In Biomedical Engineering Advances (Vol. 4, p. 100048). Elsevier BV. https://doi.org/10.1016/j.bea.2022.100048
73. Riaz, M. S. B., & Javed, A. (2021). User Interface Designing Principles for Real-time Games Strategies. *Journal of Computing & Biomedical Informatics*, *2*(01), 128-139.
74. Malik, H., Anees, T. & Mui-zzud-din BDCNet: multi-classification convolutional neural network model for classification of COVID-19, pneumonia, and lung cancer from chest radiographs. Multimedia Systems 28, 815–829 (2022). https://doi.org/10.1007/s00530-021-00878-3
75. Ijaz, S., Khan, S. A., & Abdullah, M. (2020). A Study on Critical Success Factors (CSF's) of Software Development Process, Time and Quality. *Journal of Computing & Biomedical Informatics*, *1*(01), 1-14.
76. Munir, M. U., Khan, G. A., & Ammin, A. (2020). Generic Framework Related to Database Forensics and Security Countermeasures. *Journal of Computing & Biomedical Informatics*, *1*(01), 15-30.
77. Akbar, A., Sarwar, S., & Ahmad, M. U. (2020). A Policy Recommendation to Resolve E-Health Care Issues Through Non-Functional Requirements. *Journal of Computing & Biomedical Informatics*, *1*(01), 31-48.
78. Irfan, M., Amin, O., & Marium, A. (2020). Conducting User Research for Designing Better Software Products. *Journal of Computing & Biomedical Informatics*, *1*(01), 49-65.
79. H. Malik, M. S. Farooq, A. Khelifi, A. Abid, J. Nasir Qureshi and M. Hussain, "A Comparison of Transfer Learning Performance Versus Health Experts in Disease Diagnosis From Medical Imaging," in IEEE Access, vol. 8, pp. 139367-139386, 2020, doi: 10.1109/ACCESS.2020.3004766.
80. Iqbal, S., Ahmad, N., & Raza, A. (2020). Real Time Defect Identification of White Fabric in Textile Industry using Computer Vision. *Journal of Computing & Biomedical Informatics*, *1*(01), 66-79.