Optimizing bioinformatics applications: a novel approach with human protein data and data mining techniques

Preeti Thareja, Rajender Singh Chhillar
Department of Computer Science and Applications, Maharshi Dayanand University, Rohtak, India

Article Info

Article history:

Received Jun 17, 2024
Revised Feb 8, 2025
Accepted Mar 15, 2025

Keywords:

5-fold cross validation
Bioinformatics
Data mining
Deep learning techniques
Machine learning techniques
Optimization models

ABSTRACT

Biomedicine plays a crucial role in medical research, particularly in optimizing techniques for disease prediction. However, selecting effective optimization methods and managing vast amounts of medical data pose significant challenges. This study introduces a novel optimization technique, integrated bioinformatics optimization model (IBOM) for disease diagnosis, incorporating data mining to efficiently store large datasets for future analysis. Various optimization algorithms, such as whale optimization algorithm (WOA), multi-verse optimization (MVO), genetic algorithm (GA), and ant colony optimization (ACO), were compared with the proposed method. The evaluation focused on metrics like accuracy, specificity, sensitivity, precision, F-score, error, receiver operating characteristic (ROC), and false positive rate (FPR) using 5-fold cross-validation. Results indicated that the 5-fold cross-validation method achieved superior performance with metrics: 98.61% accuracy, 96.59% specificity, 88.63% sensitivity, 99.30% precision, 92.31% F-score, 10.80% error, 92.61% ROC, and a 3.00% FPR. This method was found to be the most effective, achieving an accuracy of 0.92 in disease diagnosis compared to other optimization techniques.

Corresponding Author:

Preeti Thareja
Department of Computer Science and Applications, Maharshi Dayanand University
Rohtak, Haryana, India
Email: preetithareja10@gmail.com

1. INTRODUCTION

For several decades, extreme variations and enhancements have become witnesses in the biomedicine areas [1]. Therefore, the frontier and associative areas are derived from several kinds of research and theories of comprehensive medicine, biology, and life science. The focus of the task is to employ engineering and biology methods to research and resolve issues in life science, specifically in medical fields. Moreover, biomedicine is considered the significant research and origination of biomedical data, gene chips, nanotechnology, imaging technology, and new material [2], [3]. The data mining methods has been utilized in the existing research that needs the largest storage devices and high-capability analysis tools.

The abstract data features representation is carried out to obtain a reliable and precise performance. Nevertheless, human abstraction and data analysis are inappropriate for a large amount of data with high dimensional several numbers of occurrences. Further, the data growth rate is faster compared to the manual analysis. Therefore, it is difficult to translate the raw data into an understandable way in order to provide users with understandable original data. These gathered data should be utilized properly in order to help in the clinical diagnosis and allows to define of the drug effects during the experimentation of data; hence certainly, it is significant to enable the automatic data analysis technique for evaluating the high-level data.

Recently, various fields [4], including retail communication, banking, and medical diagnostics [5], [6], with appropriate data and knowledge that was often hidden. Most of the time, processing large data and extracting appropriate data from the complex task. Therefore, data mining is considered a powerful tool for managing tasks, especially in the medical field. The classification of data [7] can be utilized to identify the results of various diseases in order to determine the genetic behaviors. Further, various techniques are used in the existing methods to classify and predict the cancer patterns and compared with the tree classifiers [8].

A survey has been conducted to provide an aspect of various models presented with the utilized data mining-data mining techniques [9], [10]. There are various data mining methods [11], including support vector machine (SVM), clustering algorithms, artificial intelligence (AI), genetic algorithm (GA), neural network (NN), decision tree (DT), and naive Bayes (NB). Moreover, various studies [12] have been conducted to enhance the accuracy of the prediction model [13] utilizing particular methods or integration of two effective methods.

The main contribution of the proposed novel integrated bioinformatics optimization model (IBOM) is to apply various optimization algorithms in the bioinformatics application in order to enhance the accuracy with the appropriate experimental details and methods implemented. Moreover, the analysis of five different optimization algorithms, namely, whale optimization algorithm (WOA), GA, ant colony optimization algorithm (ACO), multi-verse optimization algorithm (MVO), and 5-fold cross-validation algorithm, has been discussed. Several performance metrics, including accuracy, precision, recall, F-score, receiver operating characteristic (ROC), error, and false positive rate (FPR), are used to assess the effectiveness of the suggested technique. The corresponding findings are then discussed.

The coordination of this article is as follows. Section 2 offers a synopsis of prior works on protein-protein interaction (PPI) prediction. Section 3 describes our suggested method for novel IBOM optimization model. The experiment findings are shown in section 4. Lastly, section 5 concludes the work, and the following section lists references.

2. LITERATURE REVIEW

Hu and Ohue [14] demonstrates SpatialPPI, a technique that forecasts PPIs by analyzing protein complexes predicted by the AlphaFold multimer using deep neural networks (DNN). By converting the atomic coordinate and computing the atomic distribution, the protein complexes maps were found. This method uses sophisticated image processing techniques to retrieve important three-dimensional structural data from protein complexes. The suggested approach predicts PPIs with encouraging findings, demonstrating the possibilities of 3D spatial rendering methods to further structural biology research.

Cao *et al.* [15] provide a preliminary information fusion-based node representation technique that uses interaction and sequencing network profiles to present protein feature information. To be more precise, protein interaction profile and protein sequence data are recorded using distance metrics. A weighted features fusion technique is used to stable the weights of the two sources of data with a weight parameter to generate an initial information matrix. The features of proteins are then represented by training a stacked autoencoder (SAE) architecture on the original data fusion matrix. Finally, downstream prediction tasks are performed using an SVM classifier. Using a 5-fold cross-validation procedure, the method attained an accuracy of 97.69% for the Homo sapiens dataset, allowing for a full assessment of its performance.

Gündüz *et al.* [16] presented GenomeNet-Architect is a neural architecture design platform that automatically optimizes deep learning models based on genomic sequence data. It adjusts the architecture's general design, including a genomics-specific search space. It also optimizes the model training technique as well as the hyperparameters of individual layers. In comparison to the top-performing deep learning baselines, GenomeNet-Architect reduced the misinterpretation rate by 19% on a viral categorization test, requiring 67% less time for prediction and 83% fewer metrics to reach similar contig-level accuracy.

Dang and Vu [17] introduce xCAPT5, a novel hybrid classifier that uses the T5-XL-UniRef50 protein large language model to generate rich amino acid embeddings from protein sequences. The heart of xCAPT5 is a multi-kernel deep convolutional neural network (CNN) that successfully captures complicated collaborative information at the small and big levels. It is merged with the XGBoost algorithm and concatenated with pooling features in deep that makes xCAPT5 to learn important vectors with little computing cost. Experimental results reveal that xCAPT5 surpasses many approaches in predicting binary PPI, outstanding at cross-validation on numerous datasets.

Ahmed *et al.* [18] build a novel method that combines several types of smart layers and NN. Focusing on the minor details of sequences of amino acids, it is anticipated to develop more accurate predictions regarding proteins and extract characteristics. The aim is to verify the novel approach effectiveness by testing it at a broad level, which increases the comprehension of biology and bioinformatics. It created a custom DNA-binding proteins (DBPs) sorting architecture and improved it to be exceptionally precise and useful. The findings suggest that this strategy is extremely effective in detecting hidden patterns in massive data sets.

Yu *et al.* [19] presented a novel gradient tree boosting (GTB)-based PPI prediction pipeline. First, the pseudo amino acid composition (PseAAC), pseudo position-specific scoring matrix (PsePSSM), reduced sequence and index-vectors (RSIV), and autocorrelation descriptor (AD) are fused to recover the initial feature vector. Second, L1-regularized logistic regression (L1-RLR) is employed to choose the best feature subset and eliminate noise and redundancy. Ultimately, the GTB-PPI model is built. Using the Homo sapiens dataset, GTB-PPI obtained 95.15% accuracy, according to five-fold cross-validation. Furthermore, GTB-PPI may be utilized for predicting independent test datasets, and the outcomes demonstrate that it can greatly increase PPI prediction accuracy.

Simsek *et al.* [20] described a hybrid-data mining - data mining-based technique has been utilized to distinguish the significant variables used for survival change and in diagnosing breast cancer. Hence the significance of variables was determined for different periods measured as one, five, and ten. Further, the parsimonious models are utilized to perform different analyses by executing one-regression analysis techniques such as metaheuristic optimization techniques, GA and least absolute shrinkage and selection operator (LASSO). Therefore, two well-known resampling sources, synthetic minority over-sampling technique (SMOTE) and random under-sampling (RUS), were employed to enhance the classification model performance. Eventually, the two data mining models, including logistic regression (LR) and artificial neural networks (ANN) were used with 10-fold cross-validation. However, still, these techniques are not applied to other cancer types.

Vougas *et al.* [21] mainly focused on unsupervised and supervised techniques utilized explicitly in prediction applications (drug response), enhancement of various techniques in applicable models, and improved model performance. Further, a silica-screening process was also used in accordance with association rule-mining for defining genes. Incorporating omics information layers such as metabolomics, interactomics, phospho-proteomics, proteomics, and meta-genomics improves the method's applicability and enhances the silico-process method. However, the silico pipelines impact to a certain level from negative and false positive outcomes.

Thakkar *et al.* [22] presented fuzzy logic and data mining methods are utilized in diabetes diagnosis; these are used to locate appropriate patterns in large datasets using an integration of various machine learning (ML) methods, statistics, and manipulations. The expert systems like data mining and fuzzy logic are used for different aspects in order to manage the uncertainties and find the hidden information. Further, the fuzzy expert system (FES) examined the information from the available data that helps to indicate linguistic concepts in medical concepts. Therefore, various tasks have been processed while dataset selection through pre-processing by employing various methods, including normalization, and standardization. Next, in the feature extraction process, an effective method including fuzzy logic and DMon different classification algorithms has been used to enhance preciseness. Eventually, the random forest (RF) method procured 99.7%, and by employing various logic, concepts are processed with low complexity and high precision with high preciseness of 96%.

Kovalchuk *et al.* [23] utilized the multiple conceptual-framework techniques to integrate data analysis and patient flow. An association of process mining methods, text, and data are utilized to identify and assess patient flow, and clinical pathways classes (CPs). Accordingly, this technique allows automatic recognition of patients' dynamics on a certain micro-level in order to execute realistic simulations and acquire macro-level features, including queuing parameters, departmental load, and patient experience. Moreover, the automatic classification and identification of CPs utilization enhance the acute coronary syndrome (ACS) patient discrete-event simulation process. However, still, data-driven solution to large datasets is difficult during implementation.

Yang *et al.* [24] developed state-of-the-art ML techniques on webserver in order to build an effective predictive technique covering crucial absorption, distribution, metabolism, excretion, and toxicity (ADMET) features for drug discovery. Hence, admetSAR-ADMET that designed with medicinal chemists that enhance lead components along with an efficient ADMET properties. However, lack of ADMET properties in the prediction on practical platform for chemical research and drug discovery is difficult.

3. METHOD

The various effective optimization model has been utilized in bioinformatics applications that have been processed with the human protein data. This dataset has been processed in different steps, including data collection, data normalization, training and testing, and employing various optimization techniques. Moreover, the main novelty of the work is the integration of five effective algorithms, including the WOA, MVO, GA, ACO, and 5-fold cross-validation.

3.1. Materials and method

The human protein dataset has been utilized in the experimentation process. Various optimization techniques are utilized and executed on the MathWorks MATLAB R2020a version 1.0 to find the optimum results from the used optimization techniques to identify the efficiency and accuracy of the proposed technique. Therefore, the proposed technique has been compared with the other techniques to determine the outcome accuracy and performance evaluation.

3.2. Data mining optimization models

The analysis of five different optimization algorithms namely; WOA, GA [19], ACO algorithm [20], MVO algorithm, and 5-fold cross validation algorithm has been utilized in the current proposed system. The benefits of using different optimization algorithms have been followed.

3.2.1. Whale optimization algorithm

The WOA is considered a meta-heuristic optimization algorithm that helps in various terms and finds out the behavior of the bubble-net hunting of humpback whales. In the current research, this algorithm has been applied to the protein dataset that is robust and simple, and completely based on the stochastic-swarm-based optimization algorithm [25]. Normally, the population-based WOA has the capability to remove local optima and procure the best global optimal solution. This algorithm helps to resolve various unconstrained and constrained optimization issues process for practical applications in the absence of structural reformation. In the scenario, the clustering issues are solved utilizing WOA with the clustering context represented with k-clusters centers. Hence, every search agent has been constructed as Xi, and the mathematical formula is as shown in (1).

$$X_i = (Z_{i1}, Z_{i2}, Z_{i3}, \dots, Z_{ik}) \quad (1)$$

Here, k is number of clusters, Z_{ij} is indicates the jth cluster center vector that denotes the i_{th} search-agent in cluster.

3.2.2. Genetic algorithm

A GA is considered a heuristic-search model used in various bioinformatics applications and medical purposes for predicting various diseases. It is utilized to analyze optimized outcomes to search for problems in predicting the various actions based on evolutionary biology and natural selection. Basically, the GA is an efficient model for searching via complex and huge datasets. Therefore, they have the capability to find an effective solution in complex environments, especially capable of finding and resolving constrained and unconstrained optimization problems [26]. The focus of the GA from evolutionary biology includes recombination, selection, inheritance, and mutation in order to resolve the issues.

The main purpose of the GA in the current research is as follows, and it completely varies from an optimization algorithm, derivative-based, and classical methods in two forms: First, GA provoking the population in every move, wherein a traditional algorithm only produces a single point at every move. Second, GA picking only the subsequent population by estimation utilizing random generators, wherein a traditional technique picks the next point using deterministic computation. These techniques are compared with the other traditional techniques [27], and finally, it shows that the GA is robust. However, sometimes, its breakdowns because of inputs and noise presence.

3.2.3. Ant colony optimization algorithm

The main aspect of using gene-expression data for predicting various diseases and personalized treatment facilities in promising areas like medicine. Therefore, different algorithms are developed to classify different diseases according to the selected gene expression, and significant gains are carried out in the disease classification preciseness [28]. Moreover, the classification algorithms are developed in various studies and perform better by utilizing a selected feature subset with the available data.

Let us suppose there are only two paths which are P1 and P2. C1 and C2 are the pheromones for the paths P1 and P2, respectively. Let there be a graph having vertex V and edges E. Firstly, the ith path has the choosing probability, given in (2).

$$P_i = {}^{C_i}\!/_{C_1 + C_2}; where\ i = 1,2 \quad (2)$$

If $C_1>C_2$, then path P_1 has a higher probability of being chosen than the path P_2. If $C_1<C_2$, Path P_2 is the better option.

The return path is determined by two factors: the length of the path taken by ant and the rate of pheromone evaporation, as discussed as follows:

– The pheromone concentration varies with the length of the path, as illustrated in (3).

$$C_i = C_i + \frac{K}{L_i} \tag{3}$$

Where L_i is the path's length and K is the path's length-dependent constant. If the path is shorter, the pheromone concentration will be increased.

– In (4) depicts the change in concentration as a function of the rate of evaporation.

$$C_i = (1 - v) * C_i \tag{4}$$

Here, parameter v ranges from 0 to 1. If v is higher, the concentration will be lower.

3.2.4. Multi-verse optimization algorithm

The multiobjective optimization algorithm (MOO) has been used for various optimization issues with multiple objectives based on the criteria or goals and these will generally evaluate various aspects of the achieved solution, hence incommensurable and partially conflicts [29]. Therefore, an unconstrained MOO issues are determined with the mathematical expression, as in (5).

$$Minimize\ (z) = f(x) = \left(f_1(x), f_2(x), \dots, f_m(x)\right)\ with\ x = (x_1, x_2, \dots, x_n) \in X \tag{5}$$

Where x_n is dimensional decision solution in vector, X is decision space, and f(x) is objective function.

3.2.5. 5-fold cross validation algorithm

Cross-validation is considered a statistical technique that evaluates the skill of ML techniques. Normally, these techniques are employed in ML to compare and pick an appropriate and effective model for a provided predictive modelling issues due to the simple implementation, easy understanding, and lower bias estimation compared to other techniques. Further, the k-fold cross-validation is a process utilized to evaluate the model on new data. Therefore, common methods and tricks are utilized in using and selecting k-values for the dataset.

Let K: {1, …, N} □ {1, …, K} be an indexing function that specifies the division to which report I is assigned via randomization. Let F(x) be the fitted function obtained after removing the Kth part of the data. The CV estimation of the error in prediction is provided in (6).

$$CV(F) = \frac{1}{N}\sum_1^N L(y_i, f(x_i) \tag{6}$$

Here, the choice of K is 5.

3.3. Implementation of the proposed approach

This section includes a detailed description of how the proposed optimization methodology was implemented, including the approaches, tools, and techniques used. Figure 1 depicts the implementation method in steps, beginning with data collection and preprocessing to ensure high-quality input for the model. The system is then trained and tested using appropriate ML techniques, followed by the use of optimization algorithms such as WOA, MVO, GA, and ACO, as well as a 5-fold CV technique to improve model performance. Finally, the best-optimized results are identified by assessing the models against important performance criteria, ensuring that the suggested approach is useful in bioinformatics applications.

3.3.1. Data collection

Initially, the cancerous protein interaction data is collected in the current research. As a result, the experimental procedure takes into account the data interactions that occur with cancer proteins. These interactions are referred to be malignant PPIs. Furthermore, the term “noncancerous protein interaction” in the paper refers to either one or both proteins that have yet to be recognized in relation to cancer. Here we processed with the “human protein data” dataset downloaded from open access repository Figshare. The positive as well as negative observations were then separated into sets for both training and testing, with 80% of the data designated for training and the remaining for testing. Further, the parameters included in the human protein dataset such as N_protein_a, N_protein_b, P_protein_a, and P_protein_b, and the name, attributes, attribute type, and the total number of data have been listed in Table 1.

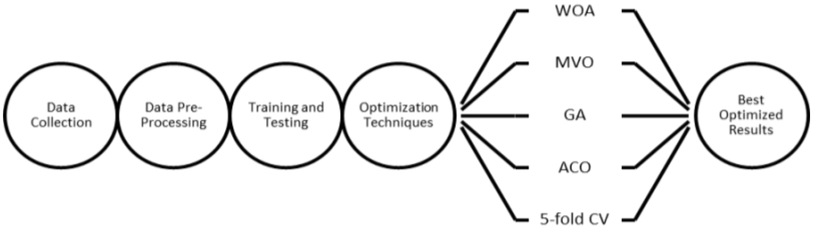

Figure 1. The implementation flow of the proposed system

Table 1. Human protein dataset

Attributes	Attribute type	Total no. of data
Negative proteins	Protein ‘a’	4262
	Protein ‘b’	4262
Positive proteins	Protein ‘a’	3899
	Protein ‘b’	3899
	Original set	16,322
	Train set	13,058
	Test set	3264

3.3.2. Data preprocessing

In this stage, we adopt a preprocessing strategy that employs two distinct measures to detect and delete interactions that are very likely to be fake. Large datasets are becoming more prevalent, and they are often challenging to comprehend. Principal component analysis (PCA) is an approach for lowering the dimension of such datasets, improving interpretability while minimizing data loss. It accomplishes this by generating new independent variables that gradually maximize variance. Finding such new variables, known as principal components, is equivalent to solve eigenvalue/eigenvector issue, and the new variables are specified by the dataset at hand rather than a priori, making PCA an adaptable data analysis technique. Weighted K-means and Gaussian mixture models (GMM) with expectation-maximization (EM) have been utilized in the research. Moreover, here we do the data normalization on the input data. As a result, normalization is frequently used to prepare data for ML. Normalization seeks to transform the numerical values of numerical columns in a dataset to a similar scale without distorting or losing information. The mathematical formulation of Weighted K-means and GMM has been defined in (7)-(11).

Weighted K-means model:

$$\phi_i(k) = \frac{exp\{-\frac{1}{\beta}||x_i - \mu_k||^2\}}{\sum_j exp\{-\frac{1}{\beta}||x_i - \mu_k||^2\}} \tag{7}$$

for k=1, …, K, and $\phi_i(k)$=>0 and $\sum_{k=1}^{K} \phi_i(k) = 1. \beta > 0$

$$\mu_k = \frac{\sum_i x_i \, \theta_i\,(k)}{\sum_i \theta_i\,(k)} \tag{8}$$

x1, x2, …, xn is data, where x ϵ Rd, μ_k is weighted average. GMM:

$$N\,(\mu, \Sigma) = \frac{1}{(2\pi)^2 \sqrt{|\Sigma|}}\, exp\,(-\,\frac{1}{2}\,(x\,-\,\mu)^T \Sigma^{-1}\,(x\,-\,\mu)) \tag{9}$$

Where μ is mean, Σ is Gaussian covariance matrix, d is number of features dataset, x is number of datapoints. GMM with EM:

– E-step:

$$\phi_i(k) = \frac{\pi_k\, N(x_i\,|\,\mu_k,\Sigma_k\,)}{\sum_j \pi_j\, N(x_i\,|\,\mu_j,\Sigma_j\,)}, \text{ for k=1, …, k} \tag{10}$$

– M-step:

$$\pi_k = \frac{n_k}{n}\,, \mu_k = \frac{1}{n_k}\sum_{i=1}^{n} \phi_i(k) x_i \;\; \Sigma_k = \frac{1}{n_k}\sum_{i=1}^{n} \phi_i(k)\,(x_i\,-\,\mu_k\,)(x_i\,-\,\mu_k)^T \tag{11}$$

for k=1, …, k, n$_k$=$\sum_{i=1}^{n} \phi_i(k)$, the values will be updated.

3.3.3. Feature extraction

In this phase, the features can be retrieved using the noise filter. Noise data is defined as the presence of oversights, duplicate information, or anomalous data in retrieved data. Feature selection is accomplished through manual evaluation of protein data. As a result, feature data are picked by manually inspecting protein data, and relevant data is obtained, reformatted, and saved in a structured database.

3.3.4. Training and testing of data

The normalized data is systematically divided into training and testing sets to facilitate further processing in the experimentation phase. In the current scenario, 80% of the data is allocated for training, ensuring that the model learns effectively from a substantial portion of the dataset. The remaining 20% is reserved for testing, allowing for an unbiased evaluation of the model’s performance and its ability to generalize to unseen data. This split is carefully chosen to maintain a balance between learning efficiency and accurate assessment, ensuring the robustness of the proposed optimization approach.

3.3.5. Optimization using the proposed model

In this step, the various optimization techniques have been utilized and these models are estimated by independent testing and cross-validation process. Further, the proposed optimization prediction model is attempting to analyze and find with the selected features then integrated with the certain specified features as a new feature of a novel prototype for further optimization. Therefore, once the secondary feature selection got over, the new optimal model was evaluated via cross-validation and independent testing. Eventually, the association amidst the best-predicted proteins utilizing an efficient deep learning techniques will be utilized to find novel therapeutic targets. Further, powerful models predict several drug-able 350 proteins that should be deeply used to find better therapeutic targets. In the current research, we have employed various optimization methods, including WOA, MVO, GA, ACO, and 5-fold cross-validation, to achieve better optimum results and the various types of optimization techniques to analyze the accuracy and performance of the proposed system. The 5-fold cross-validation procured efficient results compared to the other optimization techniques.

4. RESULTS AND DISCUSSION

4.1. Performance evaluation

The simulation results of the proposed techniques, which integrate of 5-data mining optimization algorithms-WOA, MVO, GA, ACO, and 5-fold cross-validation-have been evaluated using key performance metrics including accuracy, sensitivity, specificity, precision, and F-score. The accuracy of these optimization techniques was assessed in predicting various diseases. Comparative analysis based on performance parameters shows that 5-fold cross-validation achieved better results than the other techniques. Furthermore, the evaluation of 5-fold CV on various metrics-such as accuracy, specificity, sensitivity, precision, F-score, error rate, ROC, and FPR-demonstrated its superior effectiveness.

The performance of the proposed optimization approach was evaluated using multiple techniques, including ACO, GA, MVO, WOA, and 5-fold CV. The accuracy achieved with ACO, GA, MVO, and WOA was 0.7841, 0.5966, 0.5455, and 0.8182, respectively, whereas 5-fold CV yielded the highest accuracy of 0.9861 (98.61%). Additionally, the evaluation metrics for 5-fold CV demonstrated superior performance, with an accuracy of 98.61%, sensitivity of 88.64%, specificity of 96.59%, precision of 99.30%, F-score of 92.31%, error rate of 10.80%, ROC of 92.61%, and an FPR of 3.00%. These results highlight the effectiveness of the proposed optimization model in improving predictive performance in bioinformatics applications.

4.2. Performance comparison

A comparative analysis by comparing the performance of the proposed approach with the approach presented in the existing works has been depicted in the plot diagram. In the current proposed research, the different optimization techniques including 5-data mining optimization algorithms include the WOA, MVO, GA, ACO, and 5-fold cross-validation have experimented with certain performance evaluation metrics including accuracy, sensitivity, specificity, precision, and F-score. Eventually, Figure 2 depicts that 5-fold cross validation techniques procured better results with 0.98 accuracy compared to the other optimization techniques and denoted as a best model in diagnose of diseases.

Various optimization techniques have been executed to evaluate accuracy in the prediction of diseases. The simulation results on different performance evaluation metrics of WOA, MVO, GA, ACO, and 5-fold have been measured and achieved results on accuracy with 0.8, 0.51, 0.6, 0.68, and 0.98. Finally, compared to other optimization techniques, 5-fold has been procured with higher results of 0.98 on the accuracy, which indicates the better prediction of diseases that helps the medical experts to identify with better preciousness and generate the possibility to start the treatment soon.

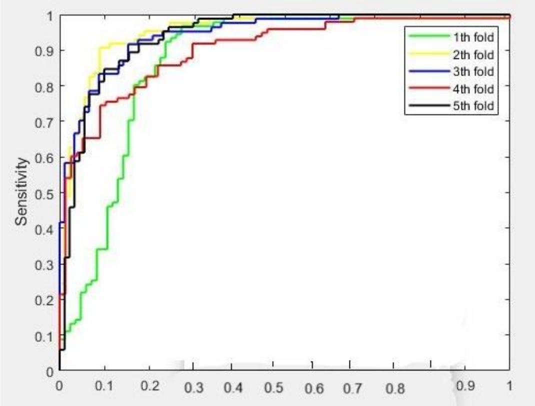

Figure 2. Results of the proposed research compared with other optimization techniques

4.3. Comparison with existing prediction models

Table 2 compares the performance of our suggested deep learning optimized model and SOTA models such as GTB-PPI, DeepPPI, MIMI-NMBAC-RF, Spatial PPI, FFANE, GenomeNet Architect, xCAPT5, and CNN-BiLG. This complete examination of classification performance includes metrics such as sensitivity, specificity, Matthews correlation coefficient (MCC), and accuracy, and the results have been validated using previous studies' findings. Our proposed model attains an excellent precision of 99.29%, which greatly excels when compared to GTB-PPI (89.99%), DeepPPI (84.32%), FFANE (98.48%), and xCAPT5 (99.1%). To begin, precision represents the minimum number of false positives. This demonstrates how effectively our deep learning algorithm can locate suitable instances of DBPs. Then, specificity is calculated as the proportion of genuine negative forecasts among all true negative instances. The model we propose excels with a specificity of 96.59% in contrast to CNN-BiLG (94.14%), Spatial PPI (79.6%), GTB-PPI (91.15%), DeepPPI (89.44%), and MIMI-NMBAC-RF (86.81%). This demonstrates how reliable our approach is at identifying negative cases, which improves the general validity of the classification results.

Table 2. Comparison with other SOTA predictors

SOTA method	Dataset used	Accuracy	Reference
Spatial PPI	Mammalian non interacting proteins pairs	0.83	[14]
FFANE	Homo sapiens	0.97	[15]
GenomeNet architect	Bacterial and viral genomes	0.83	[16]
xCAPT5	Human pan dataset	0.97	[17]
CNN-BiLG architecture	DNA binding proteins: arabidopsis and yeast	0.94	[18]
GTB-PPI	Homo sapiens	0.95	[19]
DeepPPI	Homo sapiens	0.93	[30]
MIMI+NMBAC+RF	Homo sapiens	0.94	[31]
Proposed	Homo sapiens	0.98	Current study

5. CONCLUSION AND FUTURE WORK

Biomedicals has become a significant industry in the medical platform. The main focus of the research is to develop effective optimization techniques to resolve most of the most complex issues. Nevertheless, predicting the efficient and appropriate method is challenging, and at the same time, storing a huge amount of medical data on a platform or device is complex in most scenarios. Therefore, to solve this kind of challenge, researchers utilized various optimization techniques such as cross-validation in bioinformatics applications that have been processed with the human protein data. This dataset has been executed in various steps, including data collection, data normalization, training and testing, and employing various optimization techniques. In the current research, the novelty of the work has been done with the integration of five effective algorithms, including the WOA, MVO, GA, ACO, and 5-fold cross-validation. Normally, these techniques are developed in ML to compare, and the effective technique with an appropriate model for a provided predictive modelling issues solved and made the entire process with certain benefits like simple implementation, easy understanding, and lower bias estimation compared to other techniques. Further, the 5-fold CV has been utilized in this study, and this model evaluated the new data. In this study, a novel IBOM optimization technique has been utilized in order to diagnose diseases, and data mining concepts are utilized for storing a large amount of medical data without any interruption for the further identification process. At the same time, various optimization techniques have been experimented with and compared with the proposed techniques. Finally, 5-fold cross-validation techniques procured better results with 0.98 accuracy compared to the other optimization techniques and were denoted as the best model for diagnosing diseases. In the future, the other optimization techniques will be compared with the proposed technique to determine the accuracy of predicting diseases.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Preeti Thareja	✓	✓	✓	✓	✓	✓		✓	✓	✓			✓	
Rajender Singh Chhillar		✓				✓		✓	✓	✓	✓	✓		

C	:	**C**onceptualization	I	:	**I**nvestigation	Vi	:	**Vi**sualization
M	:	**M**ethodology	R	:	**R**esources	Su	:	**Su**pervision
So	:	**So**ftware	D	:	**D**ata Curation	P	:	**P**roject administration
Va	:	**Va**lidation	O	:	Writing - **O**riginal Draft	Fu	:	**Fu**nding acquisition
Fo	:	**Fo**rmal analysis	E	:	Writing - Review & **E**diting			

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

DATA AVAILABILITY

The data that support the findings of this study are openly available on Figshare at https://figshare.com/ndownloader/files/5353900?private_link=85fd25dd6127d1bda36e. From the available datasets, only the human data was utilized for this research. The dataset can be accessed and downloaded from the provided link.

REFERENCES

[1] Y. Zhuang *et al.*, "Deep learning on graphs for multi-omics classification of COPD," *PLOS ONE*, vol. 18, no. 4, 2023, doi: 10.1371/journal.pone.0284563.

[2] Z. Gao *et al.*, "Hierarchical graph learning for protein–protein interaction," *Nature Communications*, vol. 14, no. 1, 2023, doi: 10.1038/s41467-023-36736-1.

[3] Z. Hou, Y. Yang, Z. Ma, K. chun Wong, and X. Li, "Learning the protein language of proteome-wide protein-protein binding sites via explainable ensemble deep learning," *Communications Biology*, vol. 6, no. 1, 2023, doi: 10.1038/s42003-023-04462-5.

[4] J. Levy *et al.*, "Artificial intelligence, bioinformatics, and pathology: emerging trends part I—an introduction to machine learning technologies," *Advances in Molecular Pathology*, vol. 5, no. 1, pp. e1–e24, 2022.

[5] Y. Masoudi-Sobhanzadeh and A. Masoudi-Nejad, "Synthetic repurposing of drugs against hypertension: a datamining method based on association rules and a novel discrete algorithm," *BMC Bioinformatics*, vol. 21, no. 1, pp. 1-21, 2020, doi: 10.1186/s12859-020-03644-w.

[6] T. T. Le, W. Fu, and J. H. Moore, "Scaling tree-based automated machine learning to biomedical big data with a feature set selector," *Bioinformatics*, vol. 36, no. 1, pp. 250–256, 2020, doi: 10.1093/bioinformatics/btz470.

[7] T. Tang *et al.*, "Machine learning on protein–protein interaction prediction: models, challenges and trends," *Briefings in Bioinformatics*, vol. 24, no. 2, 2023, doi: 10.1093/bib/bbad076.

[8] P. Maurya and N. P. Singh, "Mushroom classification using feature-based machine learning approach," in *Proceedings of 3rd International Conference on Computer Vision and Image Processing*, 2020, pp. 197–206, doi: 10.1007/978-981-32-9088-4_17.

[9] P. Thareja and R. S. Chhillar, "Power of deep learning models in bioinformatics," in *Innovations in Data Analytics*, 2023, pp. 535–542, doi: 10.1007/978-981-99-0550-8_42.

[10] R. J. Urbanowicz, R. S. Olson, P. Schmitt, M. Meeker, and J. H. Moore, "Benchmarking relief-based feature selection methods for bioinformatics data mining," *Journal of Biomedical Informatics*, vol. 85, pp. 168–188, Sep. 2018, doi: 10.1016/j.jbi.2018.07.015.

[11] H. Gao, C. Chen, S. Li, C. Wang, W. Zhou, and B. Yu, "Prediction of protein-protein interactions based on ensemble residual convolutional neural network," *Computers in Biology and Medicine*, vol. 152, 2023, doi: 10.1016/j.compbiomed.2022.106471.

[12] H. Luo, "Proteomic and genomic data mining with applications in plant science," *Ph.D. thesis*, Department of Bioinformatics, Wageningen University, Wageningen, Netherlands, 2023, doi: 10.18174/579767.

[13] R. Syrlybaeva and E. M. Strauch, "Deep learning of protein sequence design of protein–protein interactions," *Bioinformatics*, vol. 39, no. 1, 2023, doi: 10.1093/bioinformatics/btac733.

[14] W. Hu and M. Ohue, “SpatialPPI: three-dimensional space protein-protein interaction prediction with AlphaFold multimer,” *Computational and Structural Biotechnology Journal*, vol. 23, pp. 1214–1225, 2024, doi: 10.1016/j.csbj.2024.03.009.

[15] M. Y. Cao, S. Zainudin, and K. M. Daud, “Protein features fusion using attributed network embedding for predicting protein-protein interaction,” *BMC Genomics*, vol. 25, no. 1, 2024, doi: 10.1186/s12864-024-10361-8.

[16] H. A. Gündüz *et al.*, “Optimized model architectures for deep learning on genomic data,” *Communications Biology*, vol. 7, no. 1, 2024, doi: 10.1038/s42003-024-06161-1.

[17] T. H. Dang and T. A. Vu, “xCAPT5: protein–protein interaction prediction using deep and wide multi-kernel pooling convolutional neural networks with protein language model,” *BMC Bioinformatics*, vol. 25, no. 1, 2024, doi: 10.1186/s12859-024-05725-6.

[18] N. Y. Ahmed *et al.*, “An efficient deep learning approach for DNA-binding proteins classification from primary sequences,” *International Journal of Computational Intelligence Systems*, vol. 17, no. 1, 2024, doi: 10.1007/s44196-024-00462-3.

[19] B. Yu, C. Chen, H. Zhou, B. Liu, and Q. Ma, “GTB-PPI: predict protein–protein interactions based on L1-regularized logistic regression and gradient tree boosting,” *Genomics, Proteomics and Bioinformatics*, vol. 18, no. 5, pp. 582–592, 2020, doi: 10.1016/j.gpb.2021.01.001.

[20] S. Simsek, U. Kursuncu, E. Kibis, M. A. Abdellatif, and A. Dag, “A hybrid data mining approach for identifying the temporal effects of variables associated with breast cancer survival,” *Expert Systems with Applications*, vol. 139, 2020, doi: 10.1016/j.eswa.2019.112863.

[21] K. Vougas *et al.*, “Machine learning and data mining frameworks for predicting drug response in cancer: an overview and a novel in silico screening process based on association rule mining,” *Pharmacology and Therapeutics*, vol. 203, 2019, doi: 10.1016/j.pharmthera.2019.107395.

[22] H. Thakkar, V. Shah, H. Yagnik, and M. Shah, “Comparative anatomization of data mining and fuzzy logic techniques used in diabetes prognosis,” *Clinical eHealth*, vol. 4, pp. 12–23, 2021, doi: 10.1016/j.ceh.2020.11.001.

[23] S. V. Kovalchuk, A. A. Funkner, O. G. Metsker, and A. N. Yakovlev, “Simulation of patient flow in multiple healthcare units using process and data mining techniques for model identification,” *Journal of Biomedical Informatics*, vol. 82, pp. 128–142, 2018, doi: 10.1016/j.jbi.2018.05.004.

[24] H. Yang *et al.*, “AdmetSAR 2.0: Web-service for prediction and optimization of chemical ADMET properties,” *Bioinformatics*, vol. 35, no. 6, pp. 1067–1069, 2019, doi: 10.1093/bioinformatics/bty707.

[25] T. Senjyu, A. Y. Saber, T. Miyagi, K. Shimabukuro, N. Urasaki, and T. Funabashi, “Fast technique for unit commitment by genetic algorithm based on unit clustering,” *IEE Proceedings: Generation, Transmission and Distribution*, vol. 152, no. 5, pp. 705–713, 2005, doi: 10.1049/ip-gtd:20045299.

[26] M. F. Khan, F. Aadil, M. Maqsood, S. H. R. Bukhari, M. Hussain, and Y. Nam, “Moth flame clustering algorithm for internet of vehicle (MFCA-IoV),” *IEEE Access*, vol. 7, pp. 11613–11629, 2019, doi: 10.1109/ACCESS.2018.2886420.

[27] S. Mahapatra and S. S. Sahu, “ANOVA-particle swarm optimization-based feature selection and gradient boosting machine classifier for improved protein–protein interaction prediction,” *Proteins: Structure, Function and Bioinformatics*, vol. 90, no. 2, pp. 443–454, 2022, doi: 10.1002/prot.26236.

[28] P. Thareja and R. S. Chhillar, “A detailed survey on data mining based optimization schemes for bioinformatics applications,” *ECS Transactions*, vol. 107, no. 1, pp. 4689–4696, 2022, doi: 10.1149/10701.4689ecst.

[29] M. O. Arowolo, M. O. Adebiyi, A. A. Adebiyi, and O. Olugbara, “Optimized hybrid investigative based dimensionality reduction methods for malaria vector using KNN classifier,” *Journal of Big Data*, vol. 8, no. 1, 2021, doi: 10.1186/s40537-021-00415-z.

[30] X. Du, S. Sun, C. Hu, Y. Yao, Y. Yan, and Y. Zhang, “DeepPPI: Boosting prediction of protein-protein interactions with deep neural networks,” *Journal of Chemical Information and Modeling*, vol. 57, no. 6, pp. 1499–1510, 2017, doi: 10.1021/acs.jcim.7b00028.

[31] Y. Ding, J. Tang, and F. Guo, “Predicting protein-protein interactions via multivariate mutual information of protein sequences,” *BMC Bioinformatics*, vol. 17, no. 1, 2016, doi: 10.1186/s12859-016-1253-9.

BIOGRAPHIES OF AUTHORS

Preeti Thareja is a computer science research scholar at Maharshi Dayanand University in Rohtak, Haryana, India. Data mining, artificial intelligence, soft computing, deep learning is among her research interests. Over the last few years, she has published 5 journal papers, 4 conference papers, and 1 book chapter, as well as two books in the subjects of Python and soft computing. She can be contacted at email: preetithareja10@gmail.com.

Rajender Singh Chhillar is a computer science professor at Maharshi Dayanand University in Rohtak, Haryana, India. He was also the head of the Department of Computer Science, the Chairman of a board of studies, and a member of the executive and academic councils. Software engineering, software testing, software metrics, web metrics, bio metrics, data warehouse and data mining, computer networking, and software design are among his research interests. Over the last several years, he has produced over 91 journal papers and 65 conference papers, as well as two books in the subjects of software engineering and information technology. He is a director of the CMAI Asia Association in New Delhi, as well as a senior member of the IACSIT in Singapore and a member of the Computer Society of India. He can be contacted at email: r.chhillar@mdurohtak.ac.in.