Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 7.056

IJCSMC, Vol. 14, Issue. 5, May 2025, pg.48 – 86

Towards Protein Functions Prediction: An Inclusive Literature Review of Artificial Intelligent Techniques and Future Research Guidelines

WAFA ALAMEEN ALSANOUSI[1]; NOSIBA YOUSIF AHMED[1]; EMAN MOHAMMED HAMID[1]; K. MURTADA ELBASHIR[2]; JIANXIN WANG[3]; NOMAN KHAN[4]; AFNAN[4]

[1]Department of Computer Science faculty of mathematical and Computer Science University of Gezira–Wad Madani–Sudan.
[2]Al–Jouf University College of Information and Computer Sciences·
[3]Central South University School of Information Science and Engineering·
[4]Visual Analytics for Knowledge Laboratory Department of Software Sejong University Seoul– South Korea.

DOI: https://doi.org/10.47760/ijcsmc.2025.v14i05.004

Abstract:
Proteins perform critical functions, and their role is closely associated with their composition. Reliable prediction of protein function using computational techniques is becoming necessary because experimental limitations make it challenging to cover the massive rate of discovered proteins and their genetic transformation. This paper surveys the protein function prediction briefly discusses the employed literature's workflow, and determines if the machine and deep learning techniques have been widely used for similar objectives. Furthermore, the paper discusses the trends of protein datasets and their features and summarises researchers' questions using these datasets. Many features have been identified and extracted, ranging from traditional physicochemicals of amino acids and techniques for selecting features and reducing the dimensionality. Distinct from previous study manuscripts, we keep a detailed review of performance evaluation metrics and compare the employed protein function prediction methods, concluding the need for efficient, effective, and adaptable protein function prediction methods in real-world scenarios. We describe the machine learning processes and their development from elementary algorithms, such as logistic regression, to more sophisticated methods, like conventional and highly developed sequential deep neural networks. The techniques of computing the hyper-parameters adopted to improve the prediction efficiency have been discussed. Many studies reviewed the implemented machine and deep learning approaches for protein function prediction compared with other methods. The critical challenge well noted in protein function prediction is getting relevant information. Then, the approaches evaluated and provided future study prospects based on the results drawn from previous studies.
This review provides important information as well as prospects. We already presented this research as a preprint [1].

Keywords: Bioinformatics, Deep learning, Gene ontology, Machine learning, Protein function prediction.

1. Introduction

In nature, proteins are made up of twenty different amino acid molecules that play crucial functions in cells [2, 3]. Protein's functions involve cell identification, signaling, modulation, regulation, detection, catalysis of reactions, structure provision, and membrane transport [4, 5]. A protein's function is determined by its structure, which is determined by its Deoxyribonucleic Acid (DNA) [6]. The protein's function depends on the activity of a particular -protein molecule [7]. This function is frequently referred to as the protein's Molecular Function (MF) to discriminate from the prolonged functionality analysis. In a situation of an inclusive consideration of protein function, it is explained as an entity in a net of its interactions. For this comprehensive function view, several terminologies, such as a semantic function or cell function, are advanced [8]. Every single protein plays a vital function in a prolonged net of interacting molecules. Protein function describes all activities that allow the harmonization of metabolic reactions and bodily functions [9, 10].

An accurate Gene Ontology (GO) description depends on numerous protein function levels [11]. The GO term is essential to recognize that a protein's molecular or biochemical function is exhibited through the data that is structured and sequenced. Moreover, protein function can be predicted using an in silico technique [12]. The same as assumed in the study [13], there are a variety of mutually dependent protein function levels that can be divided into three GO term categories: MF, Biological Process (BP), and Cellular Component (CC). The MF is a measure of molecular activity that may be predicted using computational algorithms that arrange homologous or orthologous molecules [14, 15]. The BP explains the expansive functions succeeded by MF assembly, such as a fundamental metabolic pathway.

The indirect and direct physical Protein-Protein Interaction (PPI) can be established by genomic inference techniques found in BPs [16, 17]. The CC defines the position(s) inside a cell where the protein operates. Bioinformatics-based protein function predicting and genome annotating require the component of predicting subcellular protein localization because that can help classify drug targets [18]. This constituent can be forecast via posttranslational changes, membrane interaction, predicted signal sequences, or residue composition. UniProt (https://www.uniprot.org) and Pfam (https://pfam.xfam.org) are two of the most popular protein sequence repositories [19]. UniProt is the area of protein function family databases where the protein sequence is identified but has unidentified functionality [20]. The gap has continuously increased among the functional annotations and the number of protein sequences. There is an indication of the UniProt, i.e., protein sequences are higher today than ten years ago UniProtKB (https://www.uniprot.org/help/uniprotkb). Fortunately, the number of protein sequences was correctly classified and checked in all protein databanks, but UniProtKB/SwissProt (https://www.uniprot.org/help/uniprotkb) has only slightly increased in protein sequences.

Researchers [21, 22], have used Machine Learning (ML) techniques to extract sequence patterns more than a decade ago. Protein function using ML models has shown strong predictive efficiency, even though the actual mechanisms are not well defined [23]. The increasing number of literature in which ML techniques are used in their review papers to predict protein function is reported [24]. Deep Learning (DL) has also taken on extraordinary achievements continuing the trend in other fields [25, 26]. DL is well adapted to significant data issues, and because of the rapid evolution in computational efficiency, it is now within reach [27].

Additionally, we are expanding the literature review much beyond what was conducted in 2013 to involve sources of features and DL methods [24]. Further studies have concentrated on particular ontologies and taxonomies, such as predicting enzyme functional classes [28] and subcellular localization [29]. At the same time, this analysis is intended to use an extensive range of strategies and features compatible with various taxonomies. Various computationally sophisticated strategies for protein function prediction are displayed in **Figure 1** depending on the year of publication.

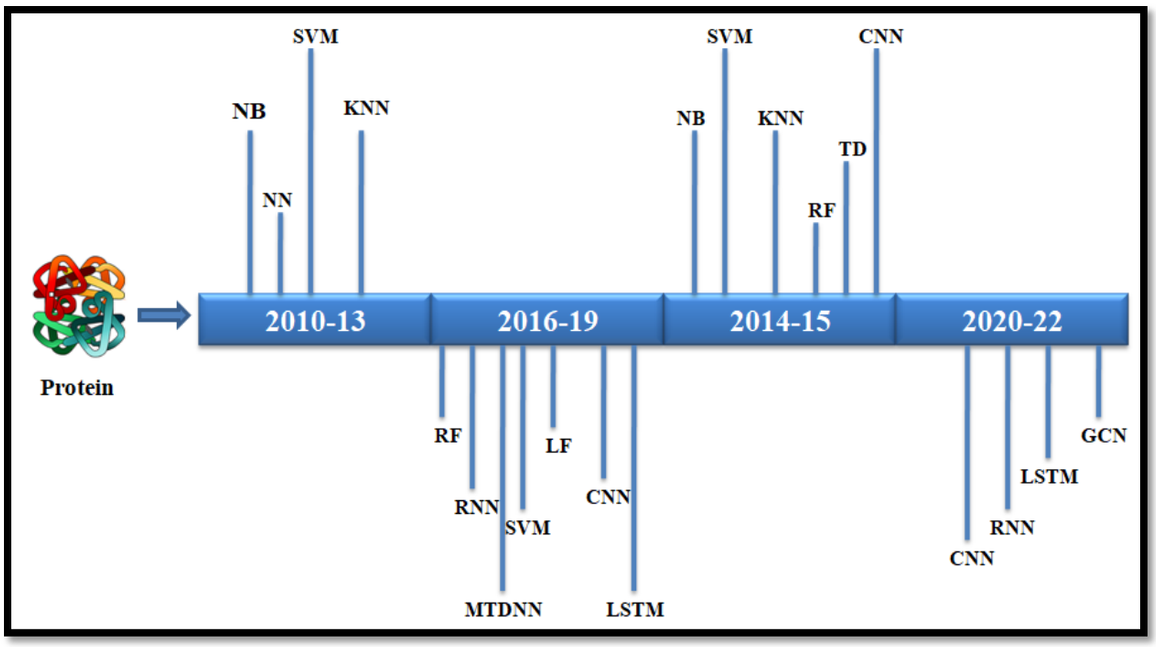


***Figure 1**: Protein function prediction methods: Computationally sophisticated strategies for protein function prediction*

The early protein function prediction trends are based on ML techniques such as Support Vector Machine (SVM), Logistic Regression (LR), Naive Bayes (NB), and K Nearest Neighbours (KNN). In the first trend (2010 - 2013), various methods utilized NB [30-32], and SVM [33-36]. The creativity of Neural Network-based methods [32, 37] and KNN-based schemes algorithms [38, 39] are also used in this trend. In the research approaches published in (2014– 2015), there are diverse methods, where some use ML (NB, KNN, etc.) and DL, such as Conventional Neural Networks (CNN) [40, 41] and hybrid methods. In (2016 – 2019), the protein function prediction methods were utilized based on the Recurrent Neural Network (RNN) [42, 43], Long Short-Term Memory(LSTM) [44, 45], and CNN. ML-based methods and hybrid techniques such as NN, SVM, and RF are also investigated in this trend. Most of the methods in (2020 – 2022) are based on CNN, RNN, Graphic Convolutional Networks (GCN) [46, 47], and Multitasking Deep Neural Networks (MTDNN) [48-50] are also used in these trends.

There are several challenges in protein function prediction. When an amino acid is mutated based on its location and type that affects the modification of protein function, such as although the amino acid is found at the active site of the enzyme, therefore, multiple mutations may have a complex effect on protein function, which is hard to predict [51]. Predicting protein function using computational tools has become necessary due to the restrictions enforced by experimental methods. Protein functions can be identified at various levels of difficulty, including physiological, molecular, phenotypic, and biochemical levels. A hierarchical structure can also be used to characterize protein function. Superoxide dismutase, for example, is a high-level oxidoreductase that converts superoxide radicals to hydrogen peroxide and molecular oxygen at a lower level [27]. Bioinformatics' critical task is to predict protein roles in BP diseases and how these roles are achieved, novel algorithms are being improved to tackle this issue [52].

It is crucial to verify these different algorithms' performance for feature prediction compared to more conventional and manual methods [53]. Initiatives, like the Critical Assessment of Function Annotation (CAFA) challenge, have tried to resolve the issue of automated protein function prediction across the Bioinformatics community [54]. This technique provides even a large-scale evaluation of statistical techniques to predict protein function [55].

One of the challenges in developing a protein function prediction method is finding a reliable dataset of multimeric proteins [56]. The data required to train researchers would come from established structural complexes in the Protein Data Bank (PDB). However, few of the structures stored in the PDB are biological multimers [57], although much work has gone into defining the three-dimensional structures of proteins. One of the many difficulties researchers face is data limitation, which can be used for training prediction methods [58]. There is no standard dataset in use due to the lack of data.

We offer an overview of the following contributions: from different viewpoints, a detailed and structured summary of ML and DL models are categorized according to the processes needed, including data preparation, feature representation, and feature selection [59, 60]. This categorization of the processes will inspire researchers who have a detailed knowledge of ML and DL preprocessing techniques [60]. Highly developed protein function prediction datasets are presented through comprehensive biological data resources such as the UniProt/SwissProt and GO annotation. Researchers may request involved biological issues and gain new technical insights, thanks to the abundance of proteomics data. This review provides the current information used in a DL-based predictive system with a combination of robust techniques efficient in predicting thousands of GO-based functional explanations [61]. It is also possible to predict the terms of GO through a meager amount of training. The protein function prediction methods, algorithms, and evaluation metrics are reviewed with novel approaches that use different ML and DL methods, from simple techniques to robust deep methods and combination methods that contribute to building strong models to help researchers in protein function prediction. We investigate the reasons that lead to the methods' effectiveness and their approaches based on their results. We also go through numerous issues that motivate researchers to build more successful algorithms.

1.1 Python code for Protein function prediction

```
from sklearn.ensemble import RandomForestClassifier
from sklearn.feature_extraction import FeatureHasher
from sklearn.pipeline import Pipeline
# Load your dataset of protein sequences and their corresponding functions
# X is a list of protein sequences, y is a list of corresponding functions
X, y = load_data()
# Define a pipeline for feature extraction and classification
pipe = Pipeline([
   ('featurizer', FeatureHasher()),
   ('classifier', RandomForestClassifier())
])
# Train the model
pipe.fit(X, y)
# Predict the function of a new protein sequence
new_sequence =
"MVHLTPEEKSAVTALWGKVNVDEVGGEALGRLLVVYPWTQRFFESFGDLSTPDAVM
GNPKVKAHGKKVLGAFSDGLAHLDNLKGTFATLSELHCDKLHVDPENFRLLGNVLVCV
LAHHFGKEFTPPVQAAYQKVVAGVANALAHKYH"
predicted_function = pipe.predict([new_sequence])
print("Predicted function:", predicted_function)
```

Explanation: This code uses a random forest classifier to predict the function of a protein sequence. The FeatureHasher is used to extract features from the protein sequences, which are then used as input to the classifier. The fit method is used to train the model on the training data, and the prediction method is used to predict the function of new protein sequences. It's important to note that this is a simple code that we build, individual may need to tweak the parameters of the pipeline and the classifier to get better results depending on the dataset are using. Also, different machine learning algorithms and deep learning architectures can be used to predict the function of proteins. These contributions include a comprehensive and in-depth analysis that varies significantly from prior review or survey work.

The rest of the paper is arranged as follows: Section 2 defines the preprocessing techniques for predicting protein function, including data preparation, feature representation, and selection. In Section 3, we summarize highly developed datasets used for predicting protein function. In Sections 4, 5, and 6, we reviewed ML and DL-based methods, algorithms implementation, and performance evaluation, respectively. Section 7 focuses on the most popular taxonomies and their results. In Section 8, we provide our conclusions and recommendations for the future.

2. Preprocessing Techniques for Protein Function Prediction

The preprocessing technique is a crucial step in ML and DL. Data consistency and the knowledge obtained from it directly impact the model's capacity to learn, so it must be organized into an appropriate form before feeding it into the model. Data preparation includes row, column, and value techniques. Feature representation techniques and algorithms are discussed after data preparation techniques. These algorithms use various features, physicochemical characteristics, amino acid structures, and amino acid structures, among the properties extracted from text. Techniques for feature extraction and dimensionality reduction are also examined.

2.1 Data Preparation Techniques

Preparing data requires converting the raw datasets into a shape that can be modeled with ML and DL methods [62]. An extensive range of methods is used to prepare data that may be used to predict models. Instead, data preparation may be viewed as a further example of hyper-parameters as components of the modeling workflow to fine-tune. Common data preparation techniques for protein function prediction are mentioned in **Table 1**. It begs what data preprocessing strategies to use, which may sound daunting for professionals and beginners. The strategy is to look at the vast area of data preparation systematically and assess data preparation techniques based on their influence on raw data regularly. Here are some common data preparation techniques for protein function prediction, along with a brief explanation of each:
It's important to note that the best data preparation technique will depend on the specific problem we are trying to solve and the types of data individuals are working with. Individuals may need to experiment with different techniques to find the one that works best for their dataset. Additionally, different techniques can be combined in a pipeline to achieve the best results.

Common data preparation techniques are needed for protein function prediction because proteins are complex molecules with many different features, such as amino acid sequence, structure, and interactions with other molecules [63, 64]. These features can vary greatly between different proteins and can have a significant impact on their function. Data preparation techniques are used to standardize and preprocess the data so that it can be used effectively by machine learning algorithms for function prediction [65, 66]. This can include steps such as data cleaning, feature extraction, and normalization, which can help to improve the accuracy and reliability of the predictions. Data preparation techniques can also help to reduce the dimensionality of the data, which can make it more manageable for machine learning algorithms [67]. This can include techniques such as feature selection, which involves identifying and removing irrelevant or redundant features from the data, and feature extraction, which involves creating new features from the existing data that are more relevant to the task at hand. Overall, the data preparation process is crucial for the success of protein function prediction, as it helps to ensure that the data is in a format that can be effectively used by machine learning algorithms and that the predictions are accurate and reliable [68, 69].

Successful data preparation necessitates coordinating and considering all possible techniques formally and thoroughly [70]. This ensures that the solution methods on a dataset are investigated and that valuable techniques are not overlooked. This may be done by combining data preparation procedures and their impact on the dataset using a framework. For example, structured ML data consists of columns, rows, and values, such as data for classification and regression stored in a Comma Separated Values (CSV) file. Preprocessing techniques that work at each stage, i.e., data preparation for values, rows, and columns, could be discussed [71]. The synthetic Minority Oversampling Technique (SMOTE) creates synthetic rows of training data for underrepresented classes [72, 73], while the random undersampling removes overrepresented classes.

Table 1: *Common data preparation techniques for protein function prediction.*

Technique	Description
One-hot encoding	Converts categorical variables into a binary matrix where each column represents a category and each row represents an observation.
Normalization	Scaling the values of the features to a common range, such as 0-1. This helps to ensure that no single feature has a disproportionate influence on the model.
Standardization	Subtracting the mean and dividing by the standard deviation of each feature

	helps standardize the distribution of the data.
Encoding amino acids	Converting amino acid sequences into numerical representations, such as one-hot encoding or using a scoring matrix like PAM or BLOSUM.
Align sequences	Aligning sequences to a reference structure or sequence can help to identify conserved regions and structural motifs.
Dimensionality reduction	Techniques such as PCA and t-SNE can be used to reduce the dimensionality of the feature space and make it easier to visualize and interpret the data.

Furthermore, different techniques for data preparation for columns include or exclude columns (features) from the data. As a result, methods like Recursive Feature Elimination (RFE) are utilized [74]. On the other hand, value preparation is a strategy for adjusting the dataset's values [75]. The most frequent preprocessing values are data transformations, which alter the volume or distribution of input parameters [76]. Data transformation, such as Gaussian, can modify data flow, reduce skews, and make it normal. Another type of data preparation, notably dimensionality reduction approaches [77]. The dimensionality reduction techniques concurrently change the columns and values [78]. Clustering algorithms aggregate techniques for adjusting the rows and values in a dataset by exchanging rows with data samples at cluster centers[78, 79]. The structure and high-level data preparation procedures are summarized in **Figure 2.**

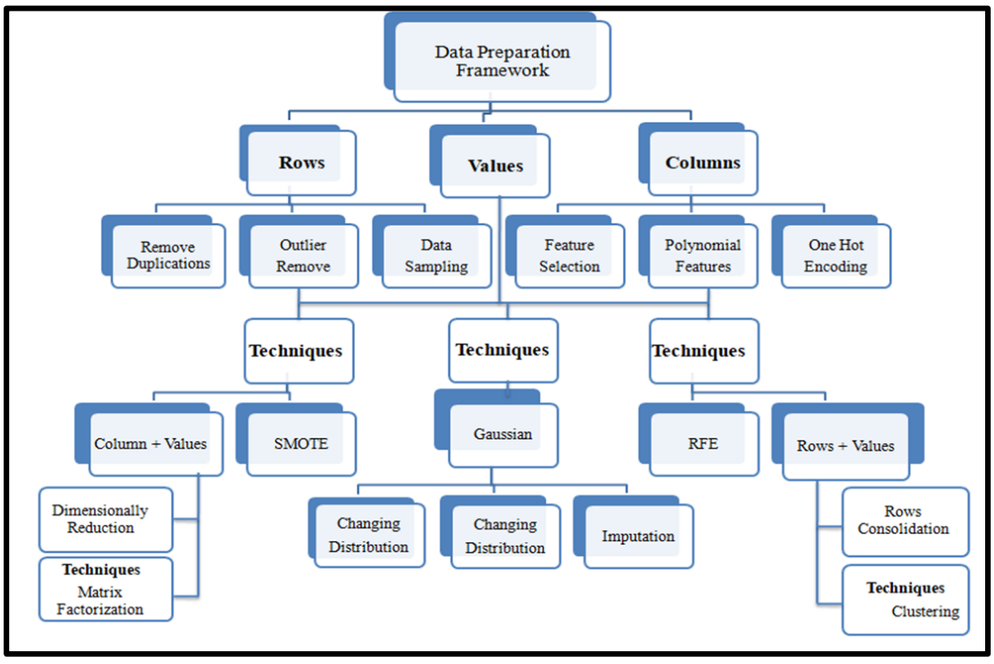


Figure 2: *Overview of different data preparation frameworks.*

2.2 Feature Representation Techniques

The features are input data to create prediction models using ML and DL methods. Identifying relevant features is crucial in applying ML or DL to any plan [80]. These features will allow the model to differentiate in a classification problem between one type of data and another or to choose an appropriate task to use other regression data analysis. A set of features representing specific objects is called a vector of features [81]. An overview of some common feature representation techniques for protein function prediction is mentioned in **Table 2.** Still, the multidimensional space connected with the features vector is known as the space of features.

Table 2: *Overview of some common feature representation techniques for protein function prediction.*

Technique	Description
One-hot encoding	Encoding a protein's amino acid sequence as a vector of binary values, where each position in the vector corresponds to a specific amino acid.
Pseudo amino acid composition	Creating a numerical vector representing the relative occurrence of various amino acids or dipeptides in a protein sequence.
Position Specific Scoring Matrix	Creating a position-specific scoring matrix (PSSM) by encoding the relative occurrence of each amino acid at each position in a multiple sequence alignment of homologous proteins.
Secondary structure prediction	Using algorithms to predict the secondary structure of a protein based on its amino acid sequence, and encoding this information as a vector of binary or numerical values.
Solvent accessibility prediction	Using algorithms to predict the solvent accessibility of each residue in a protein, and encoding this information as a binary or numerical vector.
Protein-Protein Interaction	Using experimentally determined protein-protein interactions to generate a binary matrix where an entry i, j is 1 if proteins i and j interact and 0 otherwise.

In summary, these feature representation techniques are used because they capture different types of information about proteins that can be relevant to the task of function prediction, and using a combination of these techniques can provide a more comprehensive representation of the protein, which can improve the performance of the machine learning algorithm.

Features of the protein comprise amino acid sequences, PPIs, physicochemical properties, etc. Parameters including the composition of amino acids, which relates to the presence of amino acids in a specific sequence, may be obtained from sequences or transitions of amino acids to reflect the frequency at which particular types of amino acids are complemented or anteceded with other kinds within the sequence [82]. The sequence motif is an essential category of sequence-based features that comprise amino acid sequences and is used to obtain a particular organic meaning [82]. As a result, it is essential to use specific DNA sequences as another option. Additionally, amino-terminal sequences are employed as functions. Covariance, local descriptor, and Moran autocorrelation are sequence-related characteristics proven and used for processing engagement data [83].

Protein residue physicochemical properties consist of isoelectric points, hydrophobicity, normalized volume, extinction coefficient, polarizability, polarity, molecular weight, and surface tension [84, 85]. Physicochemical interactions between proteins are mathematically represented in PPI networks [86, 87]. The linkage-based rule, also called the law of association, derives from a high functional similarity with immediate neighbor proteins and level 2 neighbors. It is also possible to predict a protein's function from its neighbors' tasks [88]. In addition to considering adjacent proteins, the interactions' weights are proportional to the experimental origins' reliability. Other network characteristics, such as an example, the average shortest path range, neighborhood connection, radiality, and topology factors, may be accessed through PPI software such as Cytoscape [89]. The overall Composition, Transformation, and Distribution (CTD) can also be used to construct features of amino acid characteristics [90].

Initially, the suggested ideas of the complexity of proteins and the probability of extracting features are mentioned [91]. Data on amino acid structure and sequence impacts are captured by protein complexity [92]. The Pseudo Amino Acid Composition (PAAC) was produced to construct a vector, for example, an amino acid sequence of any length, where ML and DL algorithms can only handle vectors [93]. If R1 R2 R3.....Rn. RL protein sequence, where R1 gives the residue at location 1, R2 gives the residue at location 2, etc., is labeled as a protein sequence with the length L, amino acid residue R1 R2 R3... RL. Dimensional vector defined by specific numbers of $20 + \lambda$. The number twenty represents the classical amino acid composition, while λ is a discrete number that demonstrates the sequence order's impact. The Position-Specific Scoring Matrix (PSSM) is typically implemented to pick out distantly associated proteins [94]. The first PSSM consists of the subsequent components; position: shows an index that has been enhanced consecutively after multiple sequence alignment of each amino acid residue in a sequence, probe: is a set of typical sequences of functionally associated proteins aligned by sequence or structural similarity, profile: A matrix of 20 columns corresponding to 20 amino acids, and consensus: is a list of residues of amino acids nearest to all probe alignment residues at each location [95]. It is created by selecting the highest score in each position in the profile [96].

Furthermore, a PSSM consists of a matrix of $W \times 20$, where W is the protein sequence length for a given protein. For the jth amino acid, the query sequence's ith position provides a score Pij with an immensely high value denoting a strongly conserved position. In contrast, a low value denotes a weakly conserved position. Nevertheless, the PSSMs need further analysis as ML algorithms usually require a fixed input size. The earlier researchers Kong and Zhang [97], performed regular analysis of three separate feature groups derived by PSSM. The average of PSSM profiles across blocks is included in the first feature group, every 5% of a sequence. Regardless of the length, the protein sequence is split into two sections, each of which has 20 features derived from the PSSMs' 20 columns. The authors concentrated on the domains with equivalent conservation rates in the second group collection rather than thinking about a sequence's domain.

The residues' physicochemical properties, which are tested using initial protein sequences, are considered in the third feature package. A total of nine physicochemical properties are classified into the mean and density groups. Hydrophobicity, isoelectric point, and mass scale are averaged, while hydrophobic, hydrophilic, polar, nonpolar, positive, and negative charge residues are used for density. As per earlier researchers, protein granularity was employed as a feature after training using ML models, including SVMs, RFs, and Decision Trees (DTs)[98, 99]. The second feature group was the most powerful in protein function prediction.

ML algorithms typically need numerical characteristics to create a fitting model [100]. For most sequences, this is simple, PPI and physiochemical features are converted into a numerical format. Using text-based features is frequently possible. Recent advancements in Natural Language Processing (NLP) methods result from broader text-based protein function prediction features from biomedical works [101, 102]. The following features obtain n-grams of amino acid sequences as text, Document to Vector (D2V), and Term Frequency-Inverse Document Frequency (TFIDF). An n-gram is a string of connected sequences from a serial dataset consisting of n objects, such as a protein sequence[103]. D2V is a dense semantic representation of documents [104]. Therefore, dense continuous vectors represent both the text and words during the training process. In TFIDF, every document is represented in a managed vocabulary by a vector of all terms. A weight is calculated as Text Frequency (TF) and Inverse Document Frequency (IDF) for each word in a document, where TF denotes the document's frequency [105]. The (IDF) is the inverse of the document. Text features are expressed in NLP via vectors and methods like Word2Vec [105]. Authors in earlier research [106, 107] show how they shaped ProtVec to describe sequences of amino acids. It has been shown that text mining applied to bioinformatics works incredibly supportive in extracting PPIs and the relationship between gene function and disease as per the investigation of earlier researchers [108-110] as investigating the PPI interactions to check the P values based reactions with a different mode of criteria’s.

Immunohistochemistry (IHC) images are also used as features in the particular case of cellular portion prediction. From September (2018) to January (2019), the Human Protein Atlas arranged a Kaggle competition to bring together biology and computer scientists to determine the locations of proteins from Immunohistochemistry images [111, 112]. Usually, feature generation needs to be driven by domain experts' measures while using classical ML algorithms and models as per investigations of earlier research [108, 113]. However, DL algorithms are capable of extracting from a given input the necessary and deep features [114]. Therefore, the generation of features is a dataset in this situation. A perfect example of data-driven feature creation is a neural network's auto-encoder, which seeks to learn its inputs [115]. In this case, the features in the center of the network are derived from the neurons' output and then trained in other classifiers. Some work has already been conducted to use auto-encoders for protein function prediction as feature generations [116, 117]. The summarizes of the most common features used to represent proteins for classification are mentioned in **Table 3**.

Table 3: *Summarizes the most common features used to represent proteins for classification.*

Feature	References	Pros	Cons
Sequence-based	[8, 13, 45, 118-140]	• The sequence is the most fundamental type of data available. • Collecting lots of data. • Good in merging with other features.	• Numeric data requires a conversation process for ML. • Suitable for structural than functional similarity. • The functional similarity is more challenging to quantify. • A significant challenge for the protein with low or no sequence similarity.
PPI networks	[48, 61-65]	• Carries rich information. • Adjacent proteins have high functional similarity.	• Limit in topological information. • Suffer from lack of reliability.
Physicochemical properties	[122-128, 141-145]	• Clear and numeric. • It can be used to predict subcellular localization.	• Providing a low amount of protein data.
Biomedical text	[104, 119, 146-150]	• Providing a rich amount of data. • Easy verification of automated predictions.	• The selection of terms affects the results strongly. • Suffer from false positive predictions.
Immunohistochemistry images	[151]	• Require more features. • Quick to visualize. • Suitable for subcellular localization tasks.	• Need larger datasets and more processing power
Representation learning	[104, 121, 129, 131, 152-154]	• Automatic feature selection and engineering.	• Needs more computational power and larger datasets.

2.2.1 Comparison of selected techniques

Duplicates can be made using the SMOTE technique by creating synthetic data points that slightly deviate from the real data points. To expand the number of instances in the dataset in a balanced way, SMOTE employs a k-nearest neighbor algorithm to randomly select data from the minority class. The k-nearest neighbor technique is used by SMOTE to randomly choose samples from the minority class and apply them. SMOTE does not use surrounding examples from other classes when creating synthetic examples because this might compromise precision and accuracy. SMOTE does not consider close instances of various classes while generating synthetic examples.

The clustering technique collects data from many sources without establishing a specific hypothesis and then utilizes clustering to identify hidden patterns in the data. To group related objects, clustering is utilized to identify similarities between various objects. NCSS Clustering includes several clustering techniques, including K-Means clustering, fuzzy clustering, and medoid partitioning. Each of these tools has undergone accuracy testing and is easy to use. You must carefully consider which clustering method would result in the best outcomes. Because different clustering procedures often provide different results, you must carefully select the clustering approach that will be most useful for your particular research project. It's possible that the labels produced by a clustering analysis won't be as effective as the data set's initial class label. Clustering makes sense if you have data but no method to classify it into meaningful

groupings. However, if your data set already has an obvious class label, the labels produced by a clustering analysis might not be as effective as the preexisting class label.

The Bayesian technique has various applications for the Gaussian Process, including regression, classification, and many more. It also allows for the use of uncertain predictions. To provide a trustworthy estimate of their uncertainty, a probability density function for these irradiance values is analytically calculated and then empirically confirmed. The expression limit of this equation suggests that only if the local standard deviation in an image is a practically stationary quantity, the distribution of image irradiance values may be deemed to be Gaussian when irradiance values are removed from an image after local mean removal.

A rapid feature selection technique such as RFE involves fitting the model, deleting the weakest features, and repeating this process until the desired number of features is attained. RFE is particularly popular because it is simple to set up and use and because it is very good at determining which features in a training dataset are most likely to be important for predicting the target variable. The RFE models might vary based on the situation at hand and the dataset. Linear regression, logistic regression, decision trees, random forests, and other common models are some that are utilized in RFE. Multiple recursive features can be used to minimize model complexity by omitting features one at a time until only the ideal number is left. Its popularity is undeniable because of how adaptable and simple to use it is, which makes it superior to other feature selection algorithms.

To create a connection between the entities of objects and users, this approach of matrix filtering is based on the idea of matrix factorization. Latent features, or the relationship between users and movie matrices, are determined to assess similarity and produce a forecast based on both item and user entities. When someone enters a search query into a search engine, the system immediately launches a matrix factorization process. After that, a suggestion is produced by the machine as a consequence of matrix factorization. The user-item interaction matrix is divided into two rectangular matrices with reduced dimensions in matrix factorization procedures. By decreasing the space's dimensions from N to K, the SVD matrix factorization approach seeks to lower the number of features that may be located in a dataset. Because this approach decreases the number of features in a dataset by decreasing the dimensions from N to K, it is more efficient in terms of time, space, and complexity.

2.3 Feature Selection Techniques

Feature Selection is the strategy used to choose the best features from a pool of candidates**[155]**. There are several positives to using this kind of feature selection, and a diverse variety of feature selection methods may be used **[156]**. There are various advantages of the feature selection process, such as improved accuracy, simple models are easier to interpret, shorter training times, enhanced generalization by reducing overfitting, easier to implement by software developers, reduced risk of data errors by model use, variable redundancy and bad learning behavior in high dimensional spaces **[157]**. The ML techniques suffer from the curse of dimensionality in many applications, such as biology this indicates that the feature space is so vast that the available information becomes scarce, resulting in output loss**[158]**. As a result, this large amount of information must be sorted out to arrive at a final list of vital features for this problem. Dimensionality reduction is the name for this stage. Some Important Feature Section Techniques based on the nature of the problem and data are mentioned in **Figure 3** and **Table 4**.

A particular case of dimensionality reduction is the selection of features, which keeps the original subset a collection of features. Each potential subset of features, which minimizes the error, should be measured and picked. The Brute Force technique is appropriate only for small sets of features. In the hypothesis space, the author in [159] considers that Feature Selection Algorithms (FSAs) may be defined as a three-factor search issue: search strategy, which is the general method for exploring the hypothesis space, candidate generation for succession and a metric of evaluation, which is the mechanism to assess the successful candidates, allows many of the hypotheses to direct the procedure for searching. At reference [160], the authors offer a broad overview of bioinformatics features' selection with particular use of these methods in sequence

alignment, microarray evaluation, and mass spectrometer. On the other hand, FSA was categorized in to stematic search, heuristic search, and hybrid approaches [161].

Table 4: *Some Important Feature Section Techniques based on the nature of the problem and data*

Sr.#	**Feature Section Techniques**	**Function and uses**
1	**Wrapper methods**	These methods use a specific machine-learning algorithm to evaluate the importance of each feature
2	**Filter methods**	These methods use statistical measures to evaluate the importance of each feature.
3	**Embedded methods**	These methods use the features selected by the machine learning algorithm itself.
4	**Lasso regularization**	This method uses L1 regularization to shrink the importance of less important features.
5	**Ridge regularization**	This method uses L2 regularization to shrink the importance of less important features.
6	**Mutual Information**	This method uses mutual information to evaluate the importance of each feature.
7	**Correlation-based feature selection (CFS)**	This method uses correlation-based feature selection to evaluate the importance of each feature
8	**Recursive Feature Elimination (RFE)**	This method recursively removes features, building models using the remaining features.
9	**Permutation Importance**	This method uses permutation to evaluate the importance of each feature by shuffling each feature and evaluating the impact on the model.
10	**Tree-based feature selection**	This method uses decision tree-based feature selection to evaluate the importance of each feature.
11	**Genetic Algorithm (GA) based feature selection**	This method uses a genetic algorithm to search for the optimal subset of features.
12	**Principal Component Analysis (PCA)**	This method uses linear algebra to find the principal components of the feature space and select a subset of features that explain the most variance.
13	**Random Forest Importance**	This method uses the feature importances from the random forest algorithm to evaluate the importance of each feature.
14	**SelectFromModel**	This method uses a trained model to select features based on their importance.
15	**Variance threshold**	This method removes all features with low variance.
16	**Chi-squared test**	This method uses a chi-squared test to evaluate the independence of each feature.
17	**t-test**	This method uses a t-test to evaluate the importance of each feature.
18	**ANOVA test**	This method uses the ANOVA test to evaluate the importance of each feature.
19	**Deep Learning Feature Selection**	This method uses deep learning models to select features based on their importance.
20	**Hybrid feature selection**	This method combines multiple feature selection techniques to achieve a better performance.

As mentioned earlier, it's important to consider the problem and data when choosing a feature selection technique. Some techniques may work better than others depending on the dataset and the problem. It's also worth noting that feature selection should be combined with feature engineering, so the best features can be used to train the model.

From a research perspective, the choice of feature representation technique for protein function prediction depends on the nature of the problem and the available data [162]. For example, if the goal is to predict the function of a protein based on its primary sequence, then feature representation techniques that capture this information, such as one-hot encoding or pseudo amino acid composition, would be more appropriate [163].

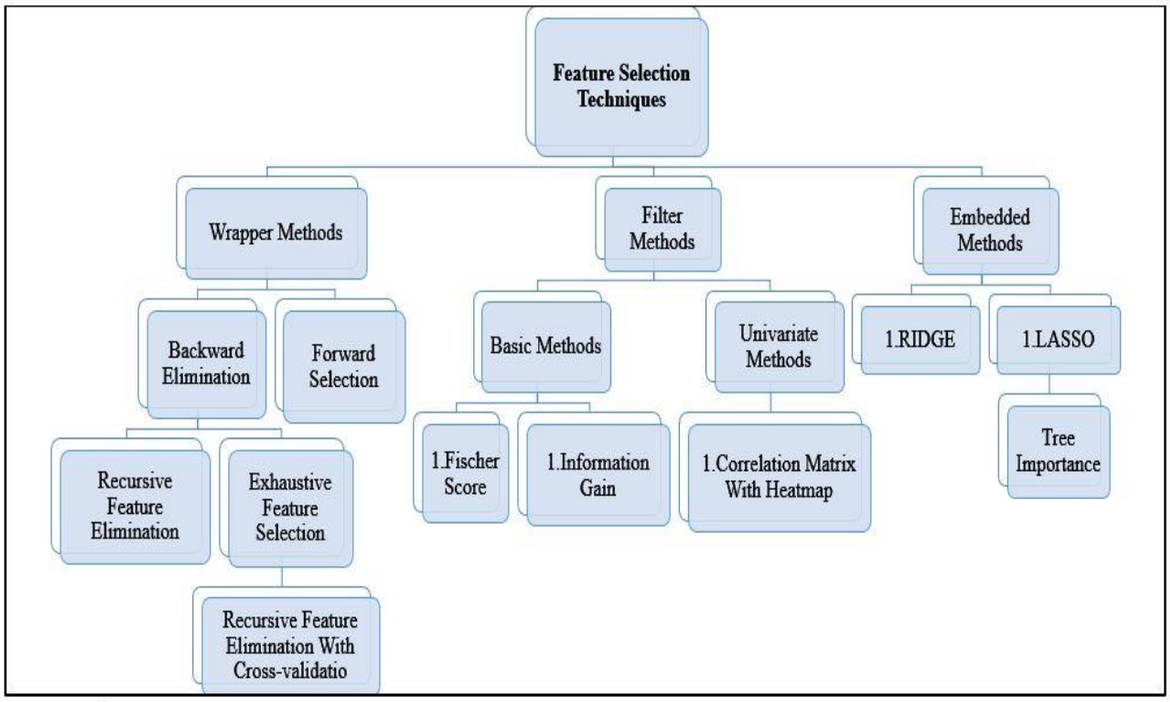


Figure 3: Comprehensive overview of Feature Selection.

If the goal is to predict the function of a protein based on its 3D structure, then feature representation techniques that capture this information, such as secondary structure prediction or solvent accessibility prediction, would be more appropriate [164]. If the goal is to predict the function of a protein based on its evolutionary conservation, then feature representation techniques such as PSSM would be more appropriate. If the goal is to predict the function of a protein based on its interactions with other proteins then feature representation techniques such as Protein-Protein Interaction would be more appropriate [165].

Furthermore, the choice of feature representation technique can also depend on the availability of data [166]. For example, if there is a large amount of experimental data available on the secondary structure or solvent accessibility of a protein, then these feature representation techniques may be more useful. On the other hand, if there is less data available, then techniques that rely more on primary sequence information may be more appropriate. It's important to note that, in practice, a combination of feature representation techniques may be used to achieve the best performance. For example, combining primary sequence information with 3D structure information can improve the accuracy of the predictions [47, 167]. In summary, the choice of feature representation technique for protein function prediction depends on the nature of the problem and the available data, and different combinations of techniques can be used to achieve the best performance [168].

2.3.1 Exploration of Feature Selection Methods

The algorithms for feature selection are commonly divided into three major categories: wrapper methods, filter methods, and embedded methods [169]. Wrapper techniques test subsets of candidate features using a prediction model with the same kind (e.g., RF or SVM), which is extended to the final classification model based on the selected features. The feature subset is used to train and evaluate the model to achieve an error rate on a hold-out set. Wrapper techniques are computationally expensive since they create a separate model for each subset but aim to perform the particular model on the best-performing feature set [170]. An example of a wrapper process is RFE. Initially, the predictive model is equipped with all possible features, and the weakest feature is eliminated until a minimum amount is reached. The authors in [171] and [172] implemented methods that used RFE.

On the other hand, the selection of forwarding features begins with assessing each particular role and choosing one as an outcome of the most effective model. Then, the selected feature and its associated features are assessed in all possible variants to choose a second feature. It replicated iteratively until it reached the maximum performance. In [173], forward feature selection is used. In [145], feature ranking is adopted using SVM as an evaluator in the WEKA

tool [174]. Instead of applying the algorithm's error rate on the selected features afterward, filter techniques assess potential feature subsets using an approximate measure. This metric is picked because it is low-cost. Famous examples include the exchange of knowledge and the Pearson product-moment correlation coefficient. Wilcoxon rank-sum, t-test, and Analysis of Variance (ANOVA) are three examples of univariate filter methods. Tang et al. [175] used the ANOVA method, called (HBPred), to rank many polypeptides for subsequent usage in training SVM classifiers to identify variations between growth hormones in DNA protein binding. To pick discriminatory characteristics, the technique depends on filters that are utilized in [8]. The Wilcoxon signed-rank test is performed for each functional class comparison feature.

Features are retained if a Wilcoxon P-value is less than 0.02 and obtained for at least one class relation. This methodology contributes potentially discriminatory results. For protein function prediction, a filter method named FrankSum was explicitly developed [176]. It combines the Wilcoxon rank with a P-value test to determine the significance level of a unique feature to distinguish between functions in groups and correlation coefficients to test feature redundancy. The Information Gain Ratio (IGR) metrics are mentioned in studies [125, 127] to rank features. XGBoost, a method that uses a gradient-boosted tree, is considered a filter technique in [81] to pick out 32 GO functions from an initial 21000 functions. The authors in [126] tested feature selection techniques for protein classification by using the rough set principle and combining correlation features, a fast correlation-based filter, and an artificial immune system. Rough sets were used in a study [177] to list the top fifteen features extracted from other sets designed based on the characteristics of the 20 amino acids' component percentages. The Minimal Redundancy Maximum Relevance (MRMR) feature selection algorithm is a maximum relevance expansion where the feature selections are most closely related to the variable of classification [178]. As biological information data also includes necessary data that is redundant, MRMR aims to resolve this issue by eliminating redundant data. For protein function prediction, many studies used MRMR [127, 135, 179].

Embedded approaches provide a wide range of strategies for extracting features throughout the model creation process. For example, the Least Absolute Shrinkage and Selection Operator (LASSO) procedure to build a linear model a regression analysis approach that combines variable selection and regularization to improve the predictability and interpretability of the final model. The (LASSO) algorithm would select any features that have non-zero coefficients. The RF, which can be used to obtain function relevance, is another example of an embedded system. In a study [180], this approach was used to rank protein sequence features to classify enzyme activity. On the other hand, a smaller set of new features is created by linearly combining the original ones by using Principal Component Analysis (PCA) or Linear Discriminant Analysis (LDA). In references [120, 141, 181, 182], PCA was used, while in [138], multi-label LDA was used. It can also be beneficial to minimize the space of potential outputs and reduce the input features' dimensionality. Two novel Label-Space Dimensionality Reduction (LSDR) techniques were developed by Makrodimitris et al. [139] to improve the CAFA efficiency of many function prediction algorithms. In the study [90], Non-Negative Matrix Factorization (NMF) was used, which can learn the semantic features of the text.

2.3.2 Comparison of Feature Selection

Experts in machine learning often use the technique of feature selection to address the high dimensionality problem by selecting a subset of the relevant qualities and excluding irrelevant, overlapping, or noisy information. Highlighting the most important critical points helps speed up and enhance the learning process. Improving feature selection procedures in machine learning is the focus of this investigation. Because different feature selection algorithms utilize different feature selection criteria, choosing the most suitable feature selection algorithm for a specific application has become more difficult as the number of feature selection techniques has increased.

Feature selection is a typical data preparation strategy for machine learning that helps narrow down the number of potential features by eliminating unnecessary ones. It enhances the effectiveness, precision, and interpretability of models constructed by learning algorithms. Genomic analysis, information retrieval, and text classification are just some of the many

domains where feature selection techniques have been used. Researchers have developed various feature selection algorithms using various criteria. However, research has shown that there is no universally good set of criteria. We created a feature selection hybrid framework based on genetic algorithms (GAs) that uses a target learning algorithm to rank features. This framework is dubbed a wrapper technique. Our framework is called HGFS, which stands for hybrid genetic feature selection.

One of the benefits of this method is that it can handle a wide variety of feature selection criteria, making it easier to zero in on feature subsets that would work well with the desired algorithm. The experimental results on genomic data show that our method is solid and efficient, able to identify subsets of features with better classification accuracy and/or less size than any individual feature selection methodology. Wrapper approaches are cumbersome and inefficient due to the frequent occurrence of multi-label classification with many characteristics in text categorization jobs. The Comparison of Feature Selection is mentioned in T**able 5.** There is no one-size-fits-all answer to which feature selection technique is the best, as the choice of technique will depend on the specific problem and dataset. However, here is a general comparison of some of the techniques mentioned above in **Table 2**: It's also worth noting that ensemble techniques, such as Hybrid feature selection, can be useful for combining the strengths of multiple feature selection techniques and achieving better performance. In general, it's always a good idea to use multiple feature selection techniques and compare the results to find the best approach for a given problem.

Table 5: *Comparison of Feature Selection.*

Sr. #	**Techniques**	**Feature**
1	**Wrapper methods & RFE**	Wrapper methods, such as Recursive Feature Elimination (RFE), are generally more computationally expensive but tend to produce more accurate results.
2	**Filter methods & CFS**	Filter methods, such as Mutual Information and Correlation-based Feature Selection (CFS), are generally less computationally expensive but may produce less accurate results.
3	**Embedded methods**	Embedded methods, such as Lasso regularization and Ridge regularization, are generally less computationally expensive and can be integrated into the training process.
4	**Tree-based feature selection**	Tree-based feature selection methods, such as Random Forest Importance and SelectFromModel, are generally effective in handling high dimensional and correlated features.
5	**PCA**	PCA is a good technique for dimensionality reduction and can be useful for dealing with high-dimensional datasets.
6	**DL models**	Feature selection based on deep learning models can be useful for datasets with complex features and patterns.

3. Datasets for Protein Function Prediction

On the way to support protein-related information management, lots of publicly reachable data depositories and sources have been created. Researchers may query complex biological issues and achievements with novel scientific understandings thanks to the abundance of proteomics data. To facilitate data-driven scientific theory creation and the finding of biological information, numerous bioinformatics databases, query capabilities, and data analysis application techniques can be applied, all that are being used to organize as well as provide biological annotations for proteins in the context of pathways, networks, and molecular genetics to support synchronous, structural, functional, and evolutionary approval. With the recent remarkable developments in genome sciences and Next-Generation Sequencing (NGS) technologies, we now have a greater understanding of the human genome [183]. **Table 6** presents a summary review of the protein function prediction datasets used in the literature. Year-wise trends of the datasets and their methods utilized for protein function prediction are shown in **Figure 4**.

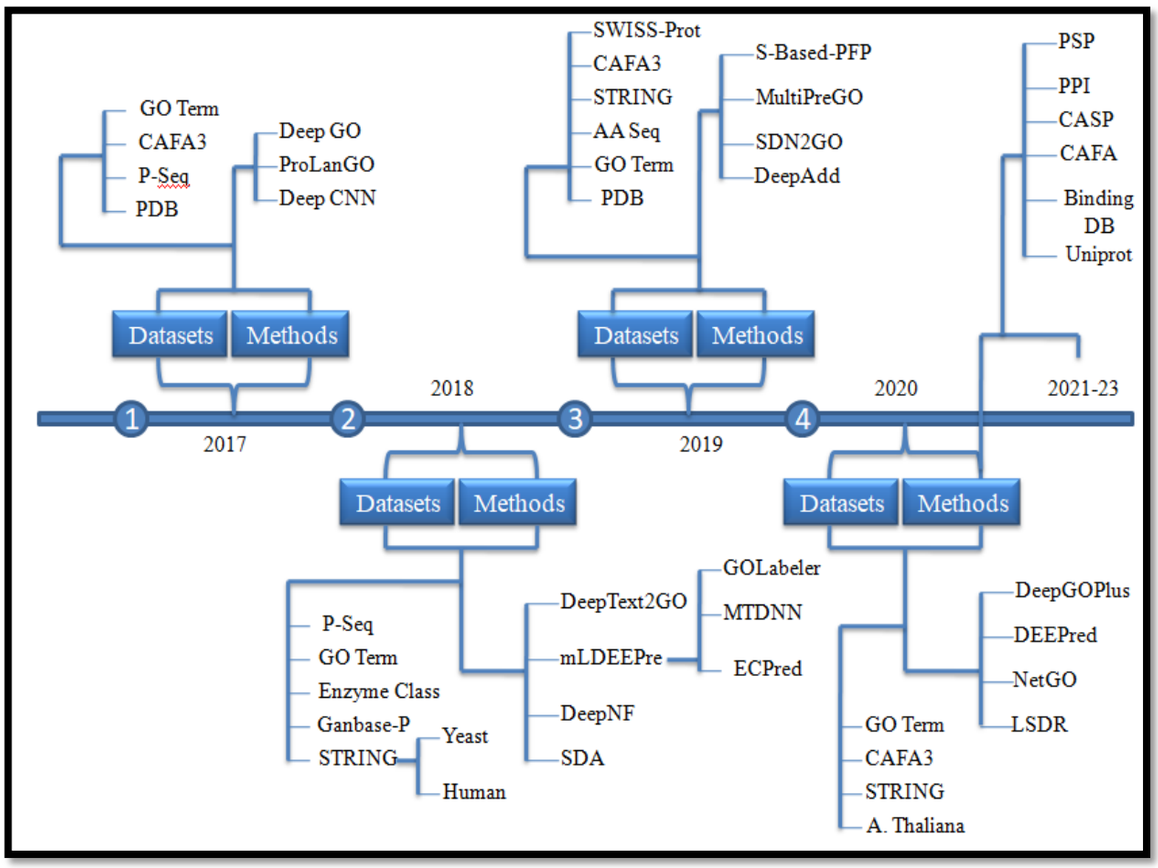


Figure 4. Year-wise trends of dataset utilized for protein function prediction.

Table 6. Summary state-of-the-art protein function prediction datasets

Dataset	References	Number of proteins	Training	Testing	Validation
GO term	[129]	BP = 19181, MF = 6221, CC = 2358	BP = 932, MF = 589, CC = 436	BP = 250, MF = 50, CC = 50	--
	[104]	BP = 50892, MF = 34008, CC = 49135	BP = 49628, MF = 33449, CC = 48153	BP = 1264, MF = 559, CC = 982	--
	[154]	BP = 7909, MF = 2545, CC = 5755	BP = 7353, MF = 2196, CC = 4906	BP = 556, MF = 349, CC = 849	--
	[184]	BP = 53150, MF = 35067, CC = 50494	BP = 51716, MF = 34488, CC = 49346	BP = 1434, MF = 679, CC = 1.148	--
UniProt-GOA dataset	[185]	932553897	--	--	--
	[186]	47658	47658	--	--
UniProt-GOA dataset for Human	[187]	13882	13882	--	--
UniProt-GOA dataset for Yeast		4796	--	4796	--
Domain data for Human (UniProtKB)		19257	19257	--	--
Domain data for Yeast (UniProtKB)		14242	--	14242	--
STRING (GO term for Yeast)	[152]	BP = 5024, MF = 4604, CC = 4549	BP = 3293, MF = 3436, CC = 3424	BP = 170, MF = 202, CC = 246	BP = 1561, MF = 966, CC = 879
STRING (GO term for Human)		BP = 13197, MF = 13360, CC = 4549	BP = 6818, MF = 8633, CC = 7656	BP = 1272, MF = 1131, CC = 1254	BP = 5107, MF = 3596, CC = 4843
STRING	[185]	9643763	--	--	--
PDB dataset	[188]	44661	35729	8932	--
	[189]	BP = 38401, MF = 42035, CC = 23297, EC = 14607	BP = 29056, MF = 31254, CC = 17396, EC = 10994	BP = 1993, MF = 2855, CC = 1535, EC = 1725	BP = 7343, MF = 7926, CC = 4366, EC = 1888
Protein sequences	[104]	147555	144842	1448422713	--
Protein sequences for Human (STRING)	[187]	19257	19257	--	--
Protein sequences for Yeast (STRING)		6507	--	6507	--
CAFA3	[42]	--	--	--	--
	[184]	BP = 55892, MF = 37247, CC = 51861	BP = 53500, MF = 36110, CC = 50596	BP = 2392, MF = 1137, CC = 1265	--
	[139]	137	--	137	--
	[186]	7173	--	BP = 3.608, MF = 1.765, CC = 1.800	--

	[190]	130787	104630	26157	--
CAFA2	[139]	860	--	860	--
Mono-functional Enzyme dataset	[146]	Single-labeled = 22168 Multi-labeled = 4076 Multi-labeled with 65% sequence similarities = 1085	Single-labeled = 15518 Multi-labeled = 2853 Multi-labeled with 65% sequence similarities = 760	Single-labeled = 6650 Multi-labeled = 1223 Multi-labeled with 65% sequence similarities = 325	--
Combination between Protein sequences and GO term	[134]	BP = 82103, MF = 47845, C = 74793	BP = 77170, MF = 45543, CC = 71388	BP = 1340, MF = 497, CC = 770, LTR1/LTR2	--
Yeast protein	[191]	2417	1934	483	--
Genbase protein		662	530	132	--
Enzyme Classes (EC)	[122]	248000	2332	24800	--
Arabidopsis thaliana (A.thaliana) proteins	[139]	7834	7834	--	--
Combination between (Amino acid sequence and 3D PDB structure)	[192]	BP = 11536, MF = 9982, CC = 10741	--	--	--
SwissProt dataset	[190]	558590	446872	111718	--
SWISS-MODEL chains	[189]	BP = 156249, MF = 190678, CC = 15046, EC = 9858	BP = 152101, MF = 149321, CC = 118118,EC = 56425	BP = 4448, MF = 4148, CC = 2947, EC = 11132	BP = 37893, MF = 37209, CC = 29395, EC = 9858

3.1 Ontology Database (GO)

The GO (https://www.geneontology.org) is a bioinformatics attempt to establish a coherent computational description of gene functions across all species at the genetic, cellular, and tissue system levels [193]. Gene products are defined in terms of BPs, CCs, and associated MFs using GO's controlled vocabulary of terms (ontologies). The use of GO terminology allows for consistent queries and associations through various biological databases. The GO website allows users to search for GO keywords, gene product annotations, and metadata throughnumerousorganisms and achieve GO enhancement studies.

3.2 Protein Sequence Database (UniProt)

The UniProt website (http://www.uniprot.org) is a free, high-quality, inclusive database of protein sequences and functional information. For morphologically heterogeneous organisms such as Archaea, Bacteria, Eukaryotes, and Viruses, the NCBI RefSeq repository offers annotated non-redundant sequences of genomic regions, transcripts, and proteins [194]. The RefSeq database is built using sequencing data from the duplicated archive database GenBank. RefSeq sequences include coding regions, conserved domains, variants, and improved annotations such as publications, names, symbols, aliases, Gene IDs, and database cross-references. A mixtureof teamwork, automatic prediction, and physicalcuration is used to create the sequences and annotations. The European Bioinformatics Institute (EBI), the Swiss Institute of Bioinformatics (SIB), and the Protein Information Resource (PIR) make up the UniProt Association. The UniProtAssociation offers a focal repository for protein sequences and functional annotations with four significant database components to assistance protein

bioinformatics studies. The UniProt Knowledgebase (UniProtKB) is the most widely used database for available information on protein sequences [195]. It contains extensive and precise annotations (protein definitions, taxonomic information, categorization, references, and citations). UniProtKB/Swiss-Prot data are humanly annotated using information from the literature and curator-evaluated computational analysis, whereas UniProtKB/TrEMBL records are computationally analyzed with automatic annotation and categorization.

3.3 3D Structure Database (PDB)

The worldwide PDB (http://www.wwpdb.org) was founded in 2003 as an international cooperation to preserve a single, freely available Protein Data Bank Archive (PDB Archive) containing macromolecular structural data [196]. Protein Data Bank in Europe (PDBe) [197], Protein Data Bank Japan (PDBj) [23], Research Collaboratory for Structural Bioinformatics Protein Data Bank (RCSB PDB) [198], and Biological Magnetic Resonance Bank (BMRB) [199] are all members of the World Wide Protein Data Bank (WWPDB).

4. Protein Function Prediction Methods and Algorithms

Many methods are used to evaluate the model's parameters based on datathat changes an input data to the desired output. Classifier models can be trained utilizingsupervised, unsupervised, and semi-supervised learning to accomplish this purpose. A training dataset is given in supervised learning with a sequence of output labels corresponding to the input vectors. In unsupervised learning, the main concern is recognizing the data's model and structure, which will necessitate more studies. Two prior frameworks make up semi-supervised learning. A combination of a small amount of labeled data and a considerableamount of unlabelled data is typically utilized in training datasets. Many additional ML and DL methods and models have rarely been utilized in the literature. The protein was represented as a document in [200], although the purpose was the protein function label. Protein sequences were grouped into a bag of words, and predictions were made using a supervised topic model. A model based on multi-label LDA trained using the protein sequence bag of words to produce features [138]. Finally, in [201] used, multi-label Gaussian kernel regression.

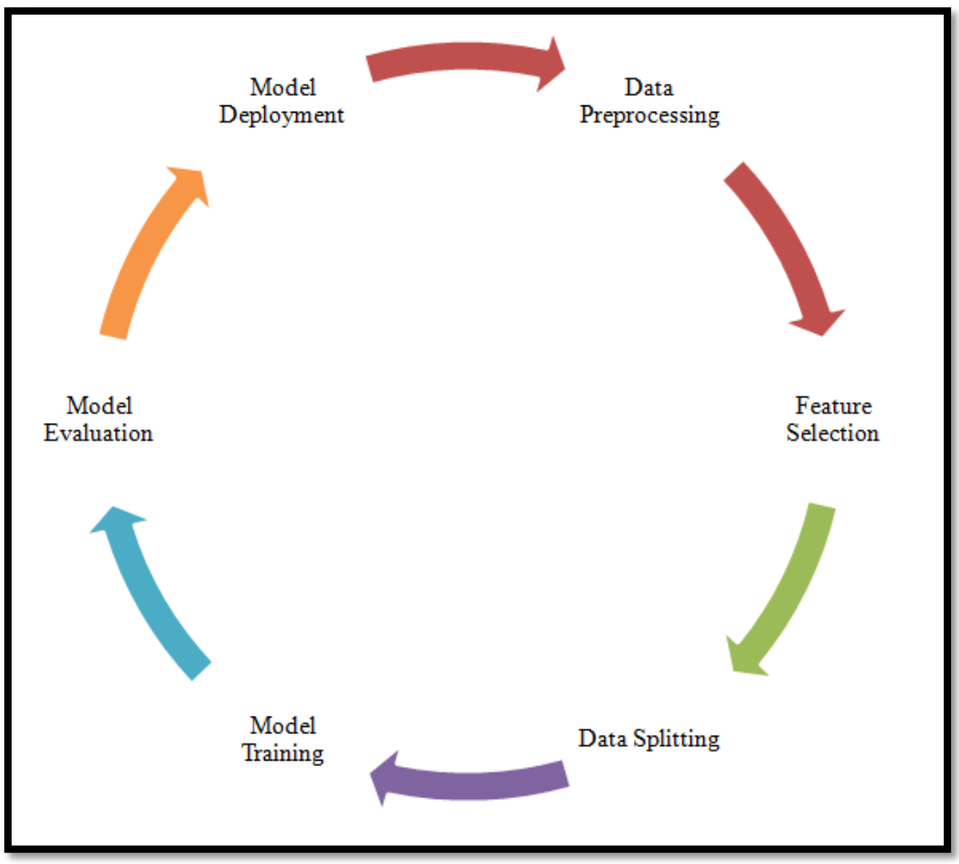


Figure 4: Protein function prediction algorithm steps

Several statistical techniques have been implemented and used in many applications in recent years. The simplest of these algorithms is the LR, which maps inputs to actual values in a range (0

and 1). LR was focused on features based on text, TFIDF, and D2V, extracted from the MEDLINE dataset of scientific publications [104]. It was completed to predict amidMF, BP, and CC. Oncematchingthe previous approach based on Markov random fields, the approach achieves better prediction accuracy. Likewise, LR was predicting protein function based on PPIs [202]. The NB classifier is usually called a simple probabilistic classifier and assumes a class feature independence. This assumption can significantly reduce the complexity of the classifier's implementation. Murakam et al. [30] used the NB classifier and Kernel Density Estimation (KDE) with sequence features (PSSM, Primordial Algorithm (PA)). The study introduces an ML approach called PSIVER for predicting PPI sites. In a study at reference [203], the authors train the NB classifier to predict the sites of PPI. At the same time, the Expanded Local Hierarchical NB algorithm [204] was used in the study at reference [205].

4.1 ML-Based Methods

Understanding protein function is essentialfor maximizing genomic data's potential in biomedical research and development. Delivering personalized health and precision medicine necessitates a comprehensive understanding of protein sequence variants' influence on phenotype. The growing distance between known proteins and their functions has prompted the creation of methods for inferring annotations automatically. Artificial intelligence and ML have a wide repertoire of algorithms and methodologies to discover and infer prediction models. The ML methods are valuableassets that could aid in discovering protein function [206]. Several ML techniques have been applied and used in various applications in the latest years.

4.1.1 K nearest Neighbours (KNN)

The KNN is a non-parametric algorithm that categorizes a certain studying a given feature space by a majority vote of the nearest k points' labels. It does not require any model training. However, the majority voting procedure suffers. In the Learning to Rank (LTR) framework, GO Labeler proposes combining five classification methods, an ML model that is excellent for multi-label classification, using features such as GO information gain, motifs, and biophysical properties sequence alignment, amino acid class label, and domains[134]. Therefore, KNN was used in references [136, 141, 148, 179, 180, 207].

4.1.2 Support Vector Machine (SVM)

SVMshave supervised learning models that process data for classification and regression analysis and come with related learning algorithms[208]. In any n-dimensional space, the SVM algorithm aims to exploit the separation between objects based on multiple classes, thus defining a maximum margin hyper-plane. Because the original function space's data is often not linearly separable, they are usually plotted on a higher-dimensional space, thus,can be easier to distinguish. Due to SVM successes in other areas, it is the most often utilized algorithm in early workswhich tried to use ML techniques to predict protein function. Moreover, SVM generally applied in[8, 118, 124, 126, 130, 135, 140, 142, 145, 149, 150, 152, 171, 173, 182, 191, 209-217].

Most of the studies use SVM to classify between types of proteins [2,63,75]. Protein - sub-cellular localization was classified with SVM in [20,70,115]. SVM is also used as a classifier with deep learning models in protein function prediction [39,110]. Other studies use SVM to classify the output of their model with a text-mining technique to predict protein function [71,123]. Web-based software for protein functional classification uses SVM for categorization [45,51]. Rough Set Theory (RST) was used for classifying protein functions using SVM, according to Abdul Rahman et al. [47]. You et al. [56] introduced a sequence-based method with SVM. Lee et al. apply high-ranked features for protein function classification using an SVM [66]. Lin et al. devised a support vector machine (SVM)-based technique for predicting mycobacterium protein subcellular localization [77]. Dos Santos et al. proposed a method for determining the classes of an enzyme dataset by introducing new variables to an SVM classifier [86]. SVM predict protein subcellular locations based on gene ontology terms, amino acid and amino acid pair composition information content[210, 216]. In [211, 212], Prediction of enzyme subfamily classes using Chou's amphiphilic pseudo-amino acid composition and a support vector machine. the application of kernel-based data fusion to protein function prediction in yeast describe in [213]. In [120] presents FFPred3, a method for assigning Gene Ontology concepts to human protein chains. Using a feature selection approach, a polipo protein was identified [121].

4.1.3 Random Forest (RF)

Random forests also referred to as random decision forests, are an ML algorithm that belongs to the ensemble learning algorithm family. Protein function is predicted using RF, which randomly samples the features to be operated as applicant features to choose the greatest among them, divides the data in each tree node, and appliessimilar packing methods called bagging [127, 172, 180, 218, 219].

4.1.4 Decision Tree (DT)

A decision tree is a tree-like model of decisions and their potential outcomes, such as chance case output, resource costs, and utility.It's one of the most fundamental ML algorithms. Each of the tree's leaves indicates a model decision. After traversing the specific path of the tree, a decision would have been made. There are a variety of DT implementations available. For classification, the C4.5 DT [126] is extensively employed [125]. The uncertain measurement used for the optimum feature selection, a novel DT classifier improved the C4.5 approach [220]. The Clus-HMC heuristic [221] was used to select the best attributes to build the tree [222].

4.1.5 Multilayer Perception (MLP)

A neural network represents a set of connected layers known as neurons. Perceptions are also known as neurons, giving birth to multilayer perception, which refers to the standard neural network design. The input layer's total number of neurons should correspond to the input features, while the output layer's total number of neurons should correspond to the number of outputs [223, 224].One-hot encoding is a suitable approach to represent the network's output for classification issues. Encoding categorical data with one hot encoding yields a binary vector with a length equal to the category variables. The array or vector is completed with 0 values, but the definite index is given a value of 1. Via an activation function, the output of a given neuron is determined, which accepts the weighted total from the previous neuron layer as an input. Sigmoid, Tanh, and Rectified Linear Unit (ReLU) are examples of standardactivation functions. Data training aims to learn appropriate weights to produce a valid output for a given input. Most neural network designs incorporate intermediary layers known as hidden layers to learn more complicated input-output data mappings. Proteins can play several roles and specialize in specific sub-functions. From this fact, the researchers designed a single-label hierarchy using local and multi-layered visualizations [225]. The use of large output labels impedes the ML algorithm's performance, so a group of 100 neural networks has been trained to predict protein function, each with 100 outputs, instead of training a single neuron network. However, the training was hierarchical, benefiting fromthe protein function's hierarchical nature [226]. Protein function was predicted using an algorithm that mimics brain function in neurobehavioral by determining the distance scale corresponding to the subsequent amino acids [227, 228]. Neural networks consider the probability density function for each class and apply Bayes' optimum decision rule for classification. The salient nonparametric estimator may be used to calculate it. [48] They utilized probabilistic neural networks to detect protein functions, outperforming KNN and SVM. Table **7** shows a summary of state-of-the-art ML methods**.**

Table 7*. Summary of state-of-the-art ML methods.*

Ref	Year	Method	Brief Description
[220]	2006	DT	This study uses priority-based packages of SDF (Sequence Derived Features) to create a new DT induction methodology.
[8]	2007	SVM + WEKA ML	This paper demonstrates that proteins with known and unknown roles vary significantly and demonstrate proteins from various bacterial species.
[228]	2011	NRProF	NRProF algorithm uses a distance metric that corresponds to the sub-sequences similarity and represents how the human brain can differentiate different sequences.
[137]	2014	RF + NB	This study predicts DNA-binding proteins using the forward best-first search technique.
[226]	2015	Multi-layer perceptron (MLP)	This research builds to explore the idea that proposes a hierarchical neural network of big-bang feed-forward.

[229]	2015	Underdamping, SMOTE, and Weighted (SVM)	This paper compares the three management strategies for protein function prediction of the imbalance problem using a Weighted SVM and SMOTE.
[219]	2016	De-novo	This work suggests a De-novo function prediction method focused on identifying function-based biophysical features.
[130]	2016	SVM-Prot	The SVM-Prot web-server uses an ML approach in this analysis. Regardless of similarities, for predicting functional protein families from protein sequences.
[225]	2016	HMC-LMLP	A novel hierarchical multi-label classification framework based on several neural networks is presented in this study.
[200]	2017	Multi-label Classification + NLP	In this study the protein was modelled as a document and a function label, converting sequences into a bag of words and estimating the model parameters.
[230]	2018	Neural Network + SVM	This study compares a neural network model to (SVM) for predicting host-pathogen PPI based on a set of characteristics.
[122]	2018	ECPred	ECPred is an automatic EC number-based enzymatic number. it adopts a supervised ensemble classification method by integrating three separate predictors from sequence data
[191]	2018	Binary-relevance (SVM)	This study proposes a DL system that builds on a stacked denoising auto-encoder that extracts powerful features to enhance prediction efficiency.
[134]	2018	GOLabeler(BLAST KNN +NB + LR)	This approach secret extracts homology data and deep-rooted data from sequence inputs and effectively and efficiently incorporates them into a model.
[104]	2018	DeepText2GO (LR + BLAST KNN)	For biomedical literature, text-based characteristics were utilized to predict protein function and deep semantic text representation was employed.
[231]	2018	XGBoost +LR+ SVM	This research extracts 21000 proteins for each human protein to fill the gap and categorizes human proteins as aging-related or non-aging-related features from various datasets.
[185]	2019	NetGO (Navie+ BLAST-KNN+LR)	This paper proposes NetGO, a web server that can improve large-scale AFP performance by incorporating massive protein-protein network information.
[139]	2019	LSDR (KNN)	Two new LSDR methods are implemented in this work, one based on the GO structure and the second based on the semantic similarity of words.
[48]	2020	SVM	This paper provides an overview of machine learning techniques used in the literature, tracing their development from basic algorithms like logistic regression to more advanced methods such as support vector machines and deep neural networks, as well as exploring hyperparameter optimization methods to enhance prediction performance.
[232]	2021	Markov Random Fields (MRF)	Predict Protein, the pioneering Internet server for protein predictions, utilized evolutionary information and ML to output multiple sequence alignments, predictions of protein structure, and predictions of protein function.
[233]	2023	Contrastive Learning–enabled Enzyme Annotation (CLEAN)	The CLEAN algorithm, an ML approach, improved the accuracy, reliability, and sensitivity of enzyme annotation compared to BLASTp by utilizing contrastive learning, enabling the annotation of understudied enzymes, correction of mislabeled enzymes, and identification of promiscuous enzymes with multiple EC numbers.

4.2 Ensembles learning

To get a better prediction, ensemble techniques integrate many base models. There are two types of ensemble techniques. Boosting is a technique used in sequential approaches to progressively create an ensemble by training each model on similar data but changing the data weights based on the last prediction error [234]. AdaBoost and Gradient Boosting are examples of such methods [235, 236]. The XGBoost method [136] is a flexible-boosting framework utilized to discover human proteins associated with aging or non-aging in the study at reference [231].

4.3 DL-Based Methods

DL is mainly involved with mastering statistics representations and feature learning, instead of being restricted to forecasting outputs with a continuous value or discrete. It enables the model to find the appropriate features automatically, ignoring conventional engineering and selecting features [237]. DL, which has many hidden layers, is typically associated with neural network architectures. Training of DL models for many activities is within reach due to the developments in computing capacity given by Graphical Processing Units (GPUs). To estimate the vast number of DNN parameters accurately, this applies only to cases where a massive amount of data is available [238]. Table **8** shows a summary of state-of-the-art DL methods.

4.3.1 Convolutional Neural Network (CNN)

CNN learns spatial hierarchies of features automatically and adaptively by back-propagation using various building blocks such as the convolution layer, pooling layer, and fully connected layers [239]. SDN2GO, a new approach suggested by Cai et al., used an integrated DL-based classification model for predicting protein function [187]. Using deep CNNs to learn and extract features from sequences, protein domains, and known PPI networks, use a weight classifier to combine these features and accurately predict GO terms. Kulmanov et al. [184] present the DeepGOPlus method for predicting protein function from the sequence by merging the deep CNN architecture with sequence similarity-based predictions. Three models (one for each GO sub-ontology) have been educated on developing DL on the amino acid sequences and PPI data. The amino acid sequences were used to create trigrams, which were then transformed to density embed, while information graph embedding was created using the PPI network features. The 1D convolutional layer was then passed through the sequence features, after that, max-pooling was carried out. The max-pooling output was merged into a fully linked layer, with the PPI network features, subsequently transferred for classification,witha sigmoid activation function in organized neural network [129]. Some CNN architectures were used, including VGG16 [240], AlexNet [241], ResNet [242, 243], InceptionNet [244, 245], and MobileNet [246] that reach high efficiency [247]. Moreover, these pre-trained models can be used on unseen data. T authors in [101] applied a CNN mixed with an SVM and a KNN classifier to predict protein structure. CNNs were trained to classify efflux protein families in transporters based on characteristics extracted in [248].

Furthermore, the DeepAdd method was developed to identify the function of a protein based on its sequence using CNN architecture [190]. CNN is also used in [147]. They propose a Multi-label Hierarchical Classification (MHC-CNN), a global method algorithm for hierarchical classification that uses a Competitive Neural Network.

***Table 8**. Summary of state-of-the-art DL methods.*

Ref	Method	Year	Brief description
[249]	DanQ framework	2016	This research aims to use a novel hybrid convolutional and bi-directional LSTM/RNN architecture to predict the non-coding function from a sequence.
[42]	ProLan (RNN)	2017	ProLan method suggests using the new proposed protein sequence language to convert the protein function problem into a language translation problem based on RNN.
[129]	DeepGO	2017	DeepGO technique proposes to predict protein function from the sequence to learn features from protein sequences and a cross-species PPI network using the DL technique.
[188]	CNN	2017	A technique is present that extracts shape features from the geometry of the 3D protein, which is inserted into a deep CNN

			enzymatic feature prediction ensemble.
[45]	RNN-LSTM	2017	This study describes protein function directly from the primary sequence without sequence alignment, heuristic scoring, or feature engineering covers using artificial (LSTM/RNN).
[152]	DeepNF	2018	To obtain high-level protein features from multiple heterogeneous interaction networks, the DeepNF proposes. DeepNF is a network fusion approach based on multimodal deep auto-encoders.
[248]	DeepEfflux (2D CNN)	2018	The DeepEfflux method is used to shield cells from the extrusion of foreign chemicals based on essential motifs.
[154]	MTDNN	2018	In this study, the MTDNN method consists of upstream shared layers. Many separate modules are stacked parallel to predict the Go term.
[184]	DeepGOPlus (1 D Convolutional + max-pooling)	2019	This model presents the DeepGOPlus method for predicting protein function from a sequence by merging deep (CNN) architecture with sequence similarity-based predictions.
[146]	mLDEEPre (KNN)	2019	The deeper method can predict the roles of multi-functional enzymes. It uses a self-adapting label assignment threshold and a unique loss function depending on the link between labels.
[250]	DEEPred	2019	DEEPred is a feed-forward deep neural network multilayer stack presented as a technique for predicting protein function based on GO keywords.
[187]	SDN2GO (CNN + weighted classifier)	2020	This study proposes SDN2GO to integrate a deep-learning-based classification model to predict protein functions. CNN layer utilizes to learn and extract features from sequences, domains, and PPI networks.
[190]	DeepADD (CNN)	2020	A DeepADD (CNN) is a new method for predicting protein function from sequence. Natural language and CNN models are used to produce word embedding and learn features from sequences.
[192]	MultiPredGO (CNN)	2020	In this study, MultiPredGO introduces a novel multimodal approach that uses the CNN technique to predict protein functions that incorporate two types of information: protein sequence and secondary structure.
[189]	Structure-Based (GCN+LSTM)	2020	In this model, the structure-based (GCN+LSTM) architecture predicts the functions more reliably than CNN, which trains solely on sequence data and previous competing approaches.
[251]	Graph neural networks (GNNs)	2023	Graph neural networks (GNNs) have been used to fuse these networks and attributes but may amplify bias from noisy edges in the PPI networks and cause over-smoothing of node representations when stacked with multiple layers.
[252]	CNN	2023	ProteInfer utilized CNN networks to directly predict protein functions, such as Enzyme Commission (EC) numbers and Gene Ontology (GO) terms, from unaligned amino acid sequences, providing precise predictions that complement alignment-based methods and allowing for lightweight software interfaces and downstream analysis.

4.3.2 Multitasking Deep Neural Networks (MTDNN)

MTDN is multi-task learning that seeks to solve several different tasks simultaneously by exploiting their correlations. For the prediction of human protein structure, three deep designs were tested [154]. The first architecture included an MTDNN, comprised of concealed layers that are

shared and hidden layers that are task-specific and was developed by the authors in [154]. In protein function prediction, deep network fusion is used. In [152], a multimodal deep auto-encoder extracts features and transfers them to an SVM classifier. In [131, 191], DL was used to analyze protein sequence embedding, which was restricted to a top length of 2000. Each amino acid may be represented in the form of a dimensional vector. After that, a convolutional layer was trained on a GPU with average pooling. In a study at [153], the authors worked on the subcellular localization of the human protein and used a stacked auto-encoder. The last layer of the DL network used SVM, RFs, and SoftMax regression to create assumptions and predictions and noticed that the best objectives were obtained with the last. Protein function prediction with text-based features has also been applied in biomedical literature, and deep semantic text representation was used in [104]. With hierarchical multi-label DL, the prediction of multifunctional enzyme function was accomplished [146].

4.3.3 Recurrent Neural Networks (RNN)

RNN is a neural network in which nodes are connected in a directed graph that follows a temporal series. This enables it to behave in a temporally complex manner by using feed-forward neural networks [253, 254]. RNN was used in the sequence to predict protein function [45]. RNNs are appropriate for sequential data processing [255], as their geometries preserve the internal state. The authors suggested a state-of-the-art strategy in [103] for converting the protein function problem into a language translation problem based on RNNs by using the novel proposed protein sequence language (ProLan) to protein function language (GOLan). The LSTM network has been widely employed in a variety of fields.

4.4 Hybrid Techniques for Protein Function Prediction

The trend in protein function prediction goes to hybrid techniques and features to predict functions better because no particular approach is generally accepted as the benchmark in this field, and each method has its limitations [256]. At reference [185], the authors proposed NetGO, a web server that can further improve the large-scale AFP's performance by incorporating massive protein-protein network information. Combining more than one algorithm instead of one algorithm was used to achieve very satisfactory results. In a study at [257], the neural networks from the SVM were merged for classification via a heuristic fusion rule. In a study at [258], a group of multilevel neural networks using multi-marker learning was specially trained to predict protein function relative to its multifunctional nature [259]. In the study [258], a two-tiered structure, the training samples for each class label are grouped into k-Medoids in the first layer. The cluster files were preserved, and the playable networks were used to calculate the essential functions between one example and the mediators. Ahmet et al. [250] proposed DEEPred, a new hierarchical multi-task DL method to solve GO-based protein function prediction. Simultaneously, an iterative algorithm is applied by authors in the study [260] used the new network to predict the function of proteins. The suggested approach can capture the interdependence among functions based on proteins and interactions.

5. Protein Function Prediction Implementations And Tools

ML systems continued to evolve over the past decade to meet the multiple specialties growing needs. Scikit-Learn is the most popular Python programming framework [157, 158]. Besides that, the R package, statistic tools, and MATLAB are also commonly used. Since the DL model requires sophisticated processing, various libraries, and frameworks, including TensorFlow [160], Keras [161], Caffe [162], and PyTorch [163] have been created to train models at faster rates on GPUs and computer clusters. As a result of the enormous volume of training data provided by GO to biomedical text data, computer-intensive architectures use GPUs to train ML and DL algorithms to predict protein functionality [47, 49, 70, 105]. This work provides examples of literature works that used hardware acceleration.

6. Protein Function Prediction Evaluation Metrics

The earlier section's ML and DL models must also be tuned to achieve much more acceptable efficiency. It implies selecting appropriate hyper-parameters before training, rather than the parameters learnedin training, which aids the model's generalization and performance even with

future data. Only a few hyper-parameters, like the optimizer and learning rate (e.g., Adam [164], RMSprop [165], or stochastic gradient descent), while the remainder isgenerally related to a particular model or algorithm. In contrast, the remainder is generally exclusive to a particular method or model. The activation function of neurons in neural networks, the neurons numbers in the hidden layers, and each layerare all possible candidates. The neural network's efficiency may also be increased by increasing the number of trainings across several epochs. Even though sets of values and thumb rules are proposed, there is no specific method for choosing the suitable hyper-parameters before training. The most common method is looking for the hyper-parameter space randomly or through grid search.Many modern ML and DL frameworks also support the finding of hyper-parameters with cross-validation. In its most basic version, Cross-validation using k-fold divides the shuffled at random testing data into k groups, with a proportion of samples utilized as training data and the rest as test data in each group. Therefore, achievingachieving an average performance outcome is possible, eliminating the so-called "lucky split".

Minimizing the imbalanced dataset, which occurs when the data in each class has no equivalent size, maybe improves ML and DL classifier performance. This study [148] evaluated the efficiency of three class-balance approaches in predicting protein function, which involves under-sampling, SMOTE, and weighted SVM. The under-sampling approach removes the additional samples in the majority of cases. In contrast, SMOTE uses synthetic minority case samples of data. Weighted SVM preserves the data size in all cases and then assigns sufficient weights in training to directly boost the minority cases' performance. However, sufficient weights are assigned during training to increase the performance of the minority cases directly. These last two methods obtained better performance, while the weighted SVM was less challenging in computation. The metrics used to determine ML and DL model's effectiveness usually include precision, accuracy, recall (sensitivity), specificity, F1-score, and hamming-loss (HL), becausepredicting protein function is commonly seen as a classification challenge. Accuracy is a measure used to assess classification models.It represents the model's percentage of accurate predictions.

$$Ac = \frac{(TP+TN)}{(TP+FP+FN+TN)} \quad (1)$$

Where Acrepresent accuracy. TP shows how many true-positive instances, TN is a true-negative case, FP stands for false-positive case, and FN indicates false negatives cases [146].

F1-score is known as the precision and recall average. It manages unequal cases more effectively compared to accuracy (because predicting the majority case can improve accuracy)[146].

$$\text{F1} - \text{score} = \frac{2(pr(\tau) * rc(\tau))}{(pr(\tau) + rc(\tau))} \quad (2)$$

$$pr(\tau) = \frac{1}{\square(\tau)} \sum_{j=1}^{pr(\tau)} \frac{\sum_i 1(s(G_j,P_j) \geq \tau).I(G_j,P_j)}{\sum_i 1(s(G_j,P_j) \geq \tau)} \quad (3)$$

$$rc(\tau) = \frac{1}{N_t} \sum_{j=1}^{N_t} \frac{\sum_i 1(s(G_j,P_j) \geq \tau).I(G_j,P_j)}{\sum_i (G_j,P_j)} \quad (4)$$

Where $pr\ (\tau)$ represents precision and $rc\ (\tau)$ represents recall, which is achieved at a particular cut-off value, which was indicated as (3), (4) [104].The value (τ) isthe number of proteins with the score no smaller than τ for at least one GO term, and 1(.) is one, if the input is true otherwise, zero.

F_{max} and S_{min} are CAFA's standard measures.The F_{max} is equal to F_1-score at the highest threshold and then cooperates with the accuracy and recall curve, whereas F_1-score is gained by reducing confusion and disinformation [104].

$$F_{max} = max_\tau \left\{ \frac{2 * pr(\tau) * \text{rc}(\tau)}{pr(\tau) + \text{rc}(\tau)} \right\} \quad (5)$$

$$S_{min}\tau = \sqrt{ru(\tau)^2 - mi(\tau)^2} \quad (6)$$

Where $ru(\tau)$ and $mi(\tau)$ are two types of errors, $ru(\tau)$ represent remaining uncertainty and $mi(\tau)$ refers to misinformation. They are specified as follows:

$$ru(\tau) = \frac{1}{N_t} \sum_{j=1}^{N_t} \sum_i ic(G_i) \,.\, 1\big(s\big(G_j, P_j\big) < \tau\big) I\big(G_j, P_j\big) \quad (7)$$

$$mi(\tau) = \frac{1}{N_t} \sum_{j=1}^{N_t} \sum_i ic(G_i) \,.\, 1\big(s\big(G_j, P_j\big) \geq \tau\big).\, 1 - I\big(G_j, P_j\big) \quad (8)$$

Where $ic(G_i)$is the information content of G_i, defined as follows:

$$ic(G_i) = log_2 \frac{1}{pr(G_i)|parents\ of\ G_i\ in\ GO} \quad (9)$$

Where $pr(G_i)|\text{parents of } G_i \text{ in GO}$ is the conditional probability of G_i given its parents of the GO structure.

HL is the percentage of incorrect labels compared to the total number of labels, i.e.

$$\text{HL} = \frac{1}{|N|\cdot|L|}\sum_{i=1}^{N}\sum_{j=1}^{L}(Y_{i,j} \oplus X_{i,j}) \quad (10)$$

$$L = \bigcup_{i=1}^{N} Y_i \quad (11)$$

where N is the total number of data sample, L is the total number of available classes, $Y_{i,j}$ is the target,Y_i is the set of labels for the i^{th} data sample, $X_{i,j}$is the prediction and $\oplus$ is "exclusive or" (XOR) operator that returns zero if the target and prediction are the same and one if they are not. Because this is a loss function, the ideal value is 0, and the upper bound is 1.

The better the output, the broader the Area Under Curve (AUC), because this typically means a greater accurate positive rating with the same false-positive rate. AUPR refers to Area Under precision-recall (PR) curve,whichwill extract a comparable metric. Besides, a Receiver Operating Characteristic (ROC) curve represents the relationship between sensitivity and (1-specificity).

Table **9** liststhe widely used metrics used in the literature to assess protein function prediction classifiers' efficiency. Usually, many metrics appear to be evaluated in a given job, and most of them provide additional data. While it has been demonstrated that some models work more effectively than others when mapping inputs and outputs in different ways. Evaluation metrics are necessary to assess various domains' ML models to discover which produce the greatest results. To assess the appropriateness of utilizing dissimilarity representations, the efficiency of LR, NB, SVMs, DT, and a neural network were examined [261]. The SVM technique provides the best F1-score and AUC metrics performance. SVM was compared to extreme ML [120], while Gaussian NB was trained with polynomial method, RBF kernels, DT, RF, LR, KNN, and SVM [137].[262] used SVM and KNN algorithms to train sequence motifs for enzyme classification. Finally, the study [230] compared primary neural networks with SVM for predicting protein-protein interactions in human Bacillus anthracis.

7. ML And DL Applications for Protein Function Prediction

ML and DL algorithms can use a basic classification scheme as the ground truth to predict protein function. GO ontology [11], Functional Catalogue (FC) [263], and Enzyme Commission (EC) [264] are the most popular taxonomies. The ontologye ontology protects three domains protects three domains: CC, MF, and BP. There are 28 significant FC annotation schemes, including basic cellular transport, metabolism, and protein function control. Every one of the major branches has a treelike, hierarchical structure. The EC scheme is a hierarchical categorization system for enzymes depending on the chemical reactions that catalyze. Seven enzyme groups make up the top level, such as oxidoreductases, hydrolases, transferases, isomerases, translocases, lyases, and ligases. The Gene Ontology (GO) is a graphical representation of words describing gene product characteristics. The three GO ontologies are all represented as DAGs (Directed Acyclic Graphs), in which a GO node may have several parents in hierarchal, however, unlike the simpler tree-based hierarchy for the EC and FC that discussed above.

***Table 9**. Examples of commonly used metrics for evaluating prediction effectiveness.*

Evaluation Metrics	Brief Description
Accuracy [45, 119, 120, 131, 132, 136, 140, 141, 152, 153, 177, 180, 201, 209-211, 265-267]	It is one of the metrics used to evaluate classification models;informally, the percentage of accurate predictions indicates that the model is correct. Usage: It is the measure of all the cases. It is mainly used when all the classes are equally important.
AUC-ROC [123, 134, 171, 210]	The Receiver Operator Characteristic (ROC) curve is an evaluation metric for binary classification problems. A probability curve plots the TPR against FPR at various threshold values and separates the signal from the noise. The Area Under the Curve (AUC) measures a classifier's ability to distinguish between classes and is used to summarize the ROC curve.

	Usage: AUC-ROC is used when the higher accurate positive rating is obtained with the same false-positive rate.
AUPR [104, 134, 139]	Area Under the Precision-Recall, Precision-Recall (PR) curves, like the closely related receiver operating characteristic (ROC) curves, isa binary classification assessment tool that visualizes performance across various thresholds. And It tells you about the model's capacity to differentiate between cases (positive instances) and non-cases (negative examples). Usage: Machine learning researchers construct a PR curve by plotting precision-recall pairs of points obtained using different thresholds on probabilistic or other continuous-output classifiers. Similarly, a ROC curve is created by plotting true/false positive rate pairs obtained using different thresholds on probabilistic or other continuous-output classifiers.
F_{max}[104, 129, 134, 150, 154]	F_{max} stands for "Frequency Maximum." This is the maximum frequency range used when capturing vibration data on a spectrum. Usage: It took into account predictions with sensitivity ranging from high to low.
S_{min}[104, 134].	S_{min} Minimum semantic distance as an extra GO prediction assessment measure. Usage: It gives more realistic estimates of the performance on unseen data.
Precision [42, 45, 48, 120, 122, 123, 130, 132, 136, 140, 150, 180, 202, 266]	Precision is a statistic that measures how many correct optimisticforecasts have been made. Usage: Precision works better for some problems, such as imbalanced classification tasks. It's obtained by dividing the total number of true positives and false positives by the number of true positives.
Recall [42, 45, 118, 120, 122, 123, 130, 132, 136, 140, 144, 180, 201, 202, 266]	A recall is a statistic that measures how many correct positive predictions were produced out of all possible optimisticpredictions. (Also known as sensitivity). Usage: Recall works better for some problems, such as imbalanced classification tasks. That is used with these problems when accuracy is not a good measure for assessing model performance. Note: Precision and recall have an inverse relationship, meaning that increasing one at the expense of the other is possible.
F1-score [42, 45, 48, 122, 123, 130, 152, 268]	The F1-score is calculated by taking the harmonic mean of precision and recall. Usage: In cases where building a balanced classification model with the best performance is optimal, the F1 score may be used to combine the two measurements of accuracy and recall.
Specificity [118, 130, 132, 144, 201, 220, 266]	Specificity is the complement of sensitivity, or the real negative ratio, which summarizes how effectively the negative classes was predicted. Usage: To construct the ROC curve, the specificity measure is utilized to estimate the fraction of real negative cases accurately predicted. The areaarea measures the model's performance measures the model's performance under the ROC curve (AUC). Important when you want to cover all true negatives. It uses when you don’t want to raise false alarms.
Hamming-Loss [191]	Hamming-Loss is the percentage of incorrectly expected labels, i.e., incorrect labels' proportion to total labels. Usage: In the hamming loss, that uses to assign each equal mark weight. Meanwhile, the hamming loss is designed for multiclass.

7.1 Gene Ontology (GO) Taxonomy

Another recent study is protein function prediction using GO terms. The CAFA challenge has made a rising a lot of sequences available for predicting GO annotations. The GOLabeler ensemble technique incorporates LR, BLAST-KNN, and Naive GO term frequencies to address the LTR issue, resulting in the best performance for MF, BP, and CC compared to other CAFA3

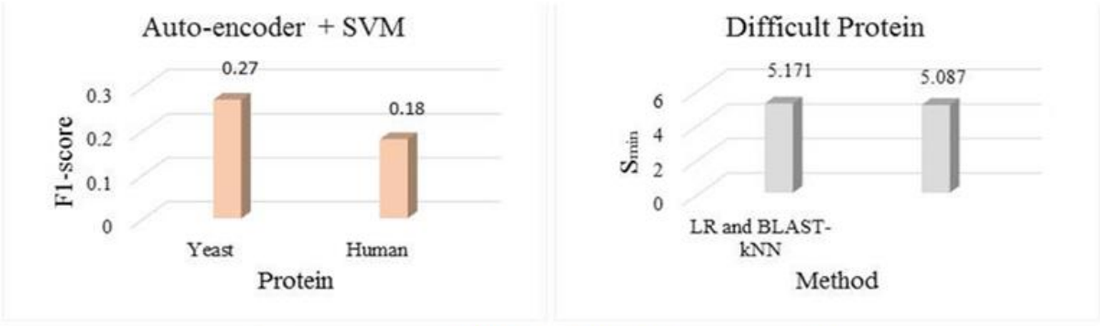
entries[134]. To predict functions related to domains, ML techniques have been used. The accuracy of prediction using ML techniques is the lowest for the BP domain, according to CAFA [54]. Many studies were trying to employ techniques which had already been proven to predict protein function in all regions. In both the MF and BP domains, a KNN classifier is utilized for prediction using text-based features collected from biomedical databases, ranging from deep neural networks [42, 129, 131, 152, 154] to KNN [139], LR [104], and SVMs [147, 214, 269].

A set of ML algorithms was developed to predict whether a particular protein would fall into one of a few categories. Using linear motifs, three different models, KNN, RF, and SVM, were trained to predict if a calmodulin-binding or mitochondrial protein was a particular protein [267]. A method to classify growth HBPred was developed [175]. SVM was also employed to detect whether a protein was an apolipoprotein. Apolipoproteins have a crucial crucialrole in the cardiovascular system and medication development. Mostcontiguous amino acid residues were utilized as input data to create dipeptides, illustrating the relation between the two and trained on them by SVMs [215]. The same ML model was used to recognize signal proteins depending on descriptors of molecular star graphs [171]. The goal is to create a binary classification model that determines which proteins are DNA-binding and non-DNA-binding. A multitude of techniques, including Gaussian NB, DTs, RFs, and LR, were used [137]. The DT classifier has been trained on the five molecular categories of the Human Protein Reference Database (HPRD), which include defending DNA, repair protein, receptor of cell surface, voltage-gated channel, and heat shock protein [220]. A KNN multi-label classifier was implemented for enzyme function prediction at the chemical process level [136]. Depending on the features extracted from the amino acid composition, rough data sets were utilized to predict the subfamilies of seven pectin lyase-like [177]. An SVM was used to predict RNA-binding, DNA-binding, and EF-hand proteins [145].

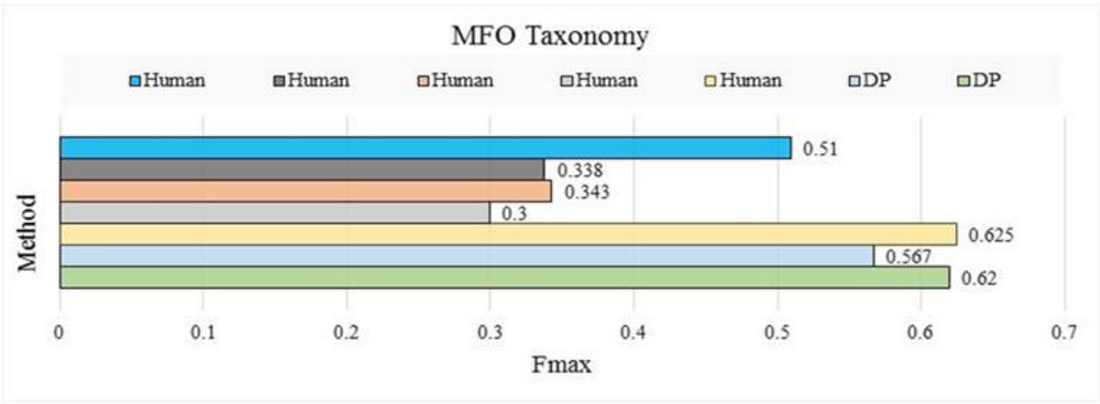
Given the large body of research that has been done on using ML methods for prediction, the problem of class imbalance in function labels has been dedicated to minimal effort. This variance is a consequence that the GO database, for example, seldom stores proteins that do not have a specific function. To decide whether a protein performs a specific role or not. The authorheauthorhe authors developed two new negative selection algorithms developed two new negative selection algorithms developed two new negative selection algorithms: Negative by Observable Bias and Negative Examples from Topic Probability [270].

The transudative multi-label ensemble classification was employed [271] to detect the functions of a protein relevant to the BP. Much of the publication applied to the CC class, which concentrated on predicting protein function relevant to a single domain. Thus in studies [175, 229, 248, 268, 272], MF was recognized. SVM was the most used ML model. It is an example in N-terminal targeting sequences, sequence motifs, amino acid composition, which are used in MultiLoc and SherLoc to predict protein subcellular localization [20, 21].Offer DeepSeq, a deep learning architecture that predicts protein activities only based on the sequence information [49]. By examining non-linear relationships among several subcellular sites, it is possible to automatically develop high-level and abstract feature representations of proteins [70]. An SVM based on n-peptide and feature selection approaches predicted a mycobacterial and Gram-negative bacteria protein [77, 115]. Also in [116, 118], an SVM was used to predict protein structure class and compositions of amino acids and amino acid pairs. Machine-learned classifiers and heterogeneous and comprehensive approaches for predicting subcellular protein localization are used in [164, 174].

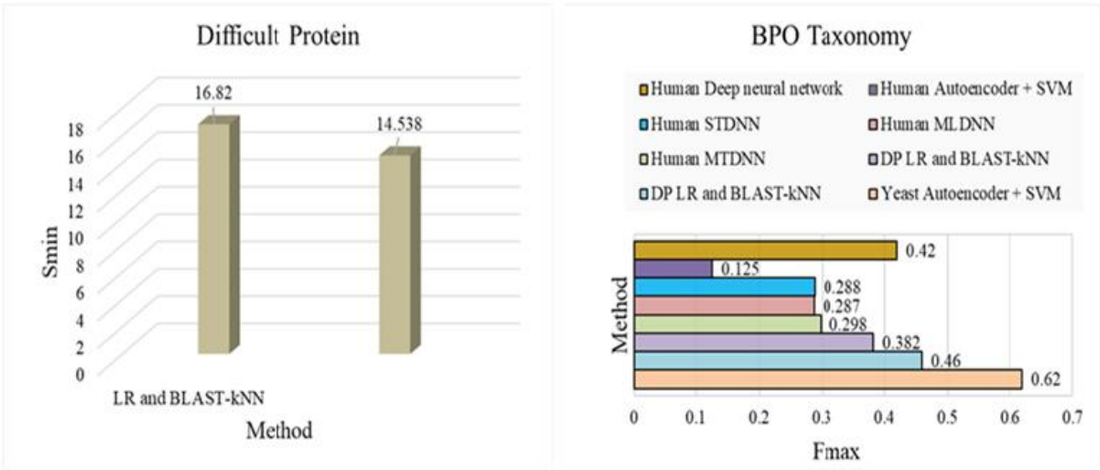
Therefore, DL and immunohistochemistry images from the Human Protein Atlas database have been used[153, 201]. An ensemble strategy was used for human protein subcellular localization [273]. While in the study [172], the authors used a RF to predict non-Golgi-resident forms of Golgi-resident proteins. Figures **4**, **5**, and **6** provide a report of the performance analysis of different ML models and algorithms for GO taxonomy of MFO, BPO, and CCO, respectively.



A **Figure 5**: A, B & C B



C



A **Figure 6:** A & B B

Figure 5. Performance comparison of various ML models and algorithms on the GO taxonomy (MF). (a) Represent the result metric (F1-score) of Auto-encoder + SVM methods with same hyper-parameter for Yeast and Human Proteins. (b) Represent the result metric (Fmax) of some methods for difficult protein (same method without hyper-parameters for two different studies) and Human Protein (with hyper-parameter) (c) Represent result metric (Smin) of method for Difficult Protein (same method without hyper-parameters for two different studies). The best performance for MFO terms was given by GoLabeler which did not make use of DL or hyper-parameter optimization. Difficult Proteins have a global sequence identity of less than 60%. **The Figure 6** show the performance comparison of various ML models and algorithms on the GO taxonomy (BP). (a) Represent the result metric (Fmax) of some methods for difficult protein (same method without hyper-parameters for two different studies), Yeast and Human Proteins (with hyper-parameter). (b) Represent result metric (Smin) of method for Difficult Protein (same method without hyper-parameters for two different studies). A similar model developed by the same authors also gave the best performance for biological process for difficult proteins. Proteins have a global sequence identity of less than 60%.

7.1 Functional Catalogue (FC) and Enzyme Commission (EC) Taxonomies

Previous work has concentrated on classification schemes such as EC and FC. The researchers trained an RF and an SVM to predict the highest level EC classification based on a sequence of amino acids, molecular weight, and chain length, among other features [123]. SVMs were used to determine each protein class precisely, with two sequences for each protein based on its structure [140]. SVMs were learned using physicochemical characteristics and sequence matching [122]. The results were outstanding. Following the FC taxonomy, protein functions were estimated through multi-label LDA [138] and employing multi-label semi-supervised graph learning associated with spatial features from PPIs [274]. Many Fun-Cat classes were predicted using the above set of input features to train a logistic regression [202]. Figure **7** provides a performance analysis report of different ML models and algorithms for the EC and FC taxonomies.

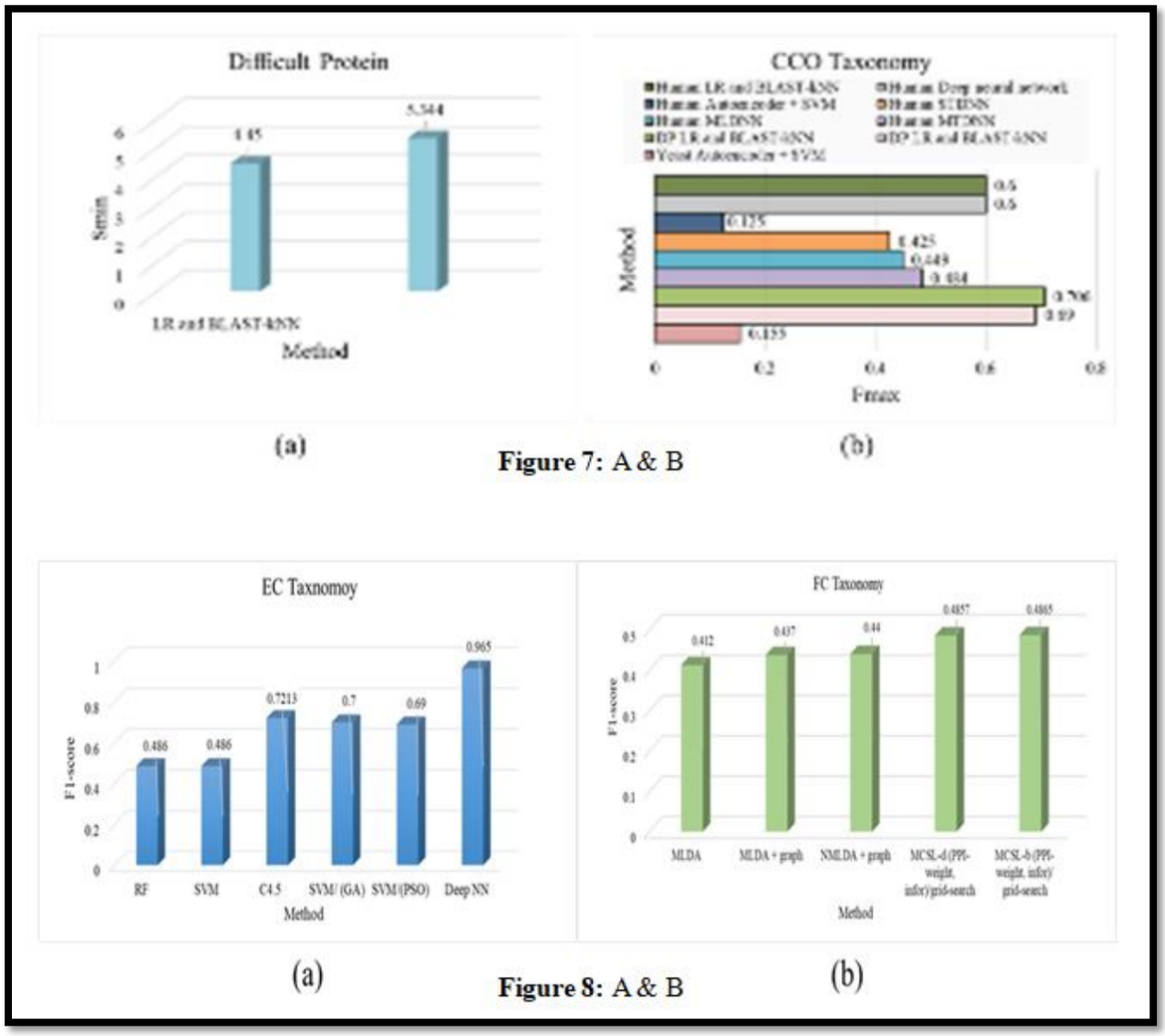


Figure 7. Performance comparison of various ML models and algorithms on the GO taxonomy (CC). (a) Represent the metric result (Fmax) of some methods for difficult protein (same method without hyper-parameters for two different studies), Yeast, and Human Proteins (with hyper-parameter). (b) Represent result metric (Smin) of the method for Difficult Protein (same method without hyper-parameters for two different studies). The CCO performed as well as a deep neural network approach (which required hyper-parameter optimization) for human proteins. **Figure 8 (a)** Performance comparison of various ML models and algorithms on the Enzyme Commission (EC) taxonomy. ML methods applied to the EC taxonomy did not disclose any optimization technique for hyper-parameters except for a grid search. Genetic algorithms and particle swarm optimization have been used in SVM, but there has not been an improvement in efficiency compared to C4.5. (b) Performance comparison of various ML models and algorithms on the Functional Catalogue (FC) taxonomy. Multi-label correlated semi-supervised learning MCSL with grid-search improved F1-score over multi-label linear discriminant analysis Multi-label LDA for the FC taxonomy.

8. Conclusion

In this survey, several predictions based on computation methods and analysis of protein function based on various biological data forms are addressed, which analyzethe evolution of ML approaches used to predict protein function based on trained data. Although there was an increase in the usage of DL strategies to extract significant functions and create good-appearing predictors, techniques using traditional ML strategies and LR still outperform DL methods. One of the challenges that DL faces is that it needs a large amount of data, possibly limiting its effectiveness, at least particularresearch on predicting protein function. Several methods in this study achieved excellent findings over a diverse range of functional groups. However, several other methods that did not produce similar outcomes need to be discussed for a variety of reasons, including the following:

- Their outcomes may improve when additional data is available for training their models.
- Technology advancements may result in improved outcomes.
- By combining these techniques with a more effective one, you may get better results than if you used them separately.

Nevertheless, scientists are now going to resort to a large amount of input features, especially those taken from biomedical texts. To close the gap between the known and unknown sequences, reliable data-driven models are essential,which would help to know the effects of protein mutations on illnesses and the development of novel proteins.

Finally, we are convinced that an effective scientific procedure can be developed in which hypotheses are produced by applying the best method for predicting functions to the scientific data that is currently available. These theories are then tested in the lab, resulting in confident predictions of a protein's function. We anticipate that the results of this survey will be helpful to computational and laboratory molecular biology professionals and complete this mission more efficiently.

List of Abbreviations

BP	Biological process	LSTM	Long short-term memory
CC	Cellular component	LR	Logistic regression
CAFA	Critical assessment of function annotation	MF	Molecular function
CNN	Convolutional neural network	ML	Machine learning
DL	Deep learning	NLP	Natural language processing
D2V	Document to vector	NB	Naive Bayes
DT	Decision tree	PPI	Protein-protein interaction
DNA	Deoxyribonucleic Acid	PSSM	position-specific scoring matrix
EC	Enzyme commission	RF	Random forest
FC	Functional catalogue	RNN	Recurrent neural networks
GO	Gene ontology	SVM	Support vector machine
KNN	K Nearest neighbors	TF	Text frequency
LDA	Linear discriminant analysis	TFIDF	Term frequency-inverse document frequency

Consent for Publication

Not applicable

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

Conflict of Interest

Declared none

Acknowledgements

None

References

1. ALSANOUSI, W.A., et al., *Towards Protein Functions Prediction: An Inclusive Literature Review of Artificial Intelligent Techniques and Future Research Guidelines.* Authorea Preprints, 2021.
2. Villegas, J.A. and E.D. Levy, *A unified statistical potential reveals that amino acid stickiness governs nonspecific recruitment of client proteins into condensates.* Protein science, 2022. **31**(7): p. e4361.
3. Toraqulova, G.B. and D.T. Mamanova, *PROTEINS IN BIOLOGY.* Scientific progress, 2022. **3**(1): p. 360-363.
4. Wu, Q., et al., *Protein arginine methylation: from enigmatic functions to therapeutic targeting.* Nature reviews Drug discovery, 2021. **20**(7): p. 509-530.
5. Karve, T.M. and A.K. Cheema, *Small changes huge impact: the role of protein posttranslational modifications in cellular homeostasis and disease.* Journal of amino acids, 2011. **2011**.
6. Eisenberg, D., et al., *Protein function in the post-genomic era.* Nature, 2000. **405**(6788): p. 823-826.
7. Yang, H., et al., *A Light‐Driven Molecular Machine Controls K+ Channel Transport and Induces Cancer Cell Apoptosis.* Angewandte Chemie International Edition, 2022: p. e202204605.
8. Al-Shahib, A., R. Breitling, and D.R. Gilbert, *Predicting protein function by machine learning on amino acid sequences–a critical evaluation.* BMC genomics, 2007. **8**(1): p. 1-10.
9. Rost, B., et al., *Automatic prediction of protein function.* Cellular and Molecular Life Sciences CMLS, 2003. **60**(12): p. 2637-2650.
10. Domán, A., et al., *Interactions of reactive sulfur species with metalloproteins.* Redox Biology, 2023: p. 102617.
11. Ashburner, M., et al., *Gene ontology: tool for the unification of biology.* Nature genetics, 2000. **25**(1): p. 25-29.
12. Friedberg, I., *Automated protein function prediction—the genomic challenge.* Briefings in bioinformatics, 2006. **7**(3): p. 225-242.
13. Lee, D., O. Redfern, and C. Orengo, *Predicting protein function from sequence and structure.* Nature reviews molecular cell biology, 2007. **8**(12): p. 995-1005.
14. DeBenedictis, E.A., et al., *Systematic molecular evolution enables robust biomolecule discovery.* Nature Methods, 2022. **19**(1): p. 55-64.
15. Guarra, F. and G. Colombo, *Computational Methods in Immunology and Vaccinology: Design and Development of Antibodies and Immunogens.* Journal of Chemical Theory and Computation, 2023. **19**(16): p. 5315-5333.
16. Zainal-Abidin, R.-A., et al., *Protein–Protein Interaction (PPI) Network of Zebrafish Oestrogen Receptors: A Bioinformatics Workflow.* Life, 2022. **12**(5): p. 650.
17. Ieremie, I., R.M. Ewing, and M. Niranjan, *TransformerGO: predicting protein–protein interactions by modelling the attention between sets of gene ontology terms.* Bioinformatics, 2022. **38**(8): p. 2269-2277.
18. Gardy, J.L. and F.S. Brinkman, *Methods for predicting bacterial protein subcellular localization.* Nature Reviews Microbiology, 2006. **4**(10): p. 741-751.
19. https://www.uniprot.org/statistics/Swiss-Prot%, U.S.-P.U.A.o. and a.o.M. 2021). *UniProt: the universal protein knowledgebase.* Nucleic acids research, 2021. **45**(D1): p. D158-D169.
20. Punta, M., et al., *The Pfam protein families database.* Nucleic acids research, 2012. **40**(D1): p. D290-D301.
21. Hossain, M.E., et al., *A systematic review of machine learning techniques for cattle identification: Datasets, methods and future directions.* Artificial Intelligence in Agriculture, 2022.
22. Karthikeyan, A. and U.D. Priyakumar, *Artificial intelligence: machine learning for chemical sciences.* Journal of Chemical Sciences, 2022. **134**: p. 1-20.
23. Xiong, G., et al., *Featurization strategies for protein–ligand interactions and their applications in scoring function development.* Wiley Interdisciplinary Reviews: Computational Molecular Science, 2022. **12**(2): p. e1567.
24. S Bernardes, J., *A review of protein function prediction under machine learning perspective.* Recent patents on biotechnology, 2013. **7**(2): p. 122-141.
25. Krishnam, N.P., et al., *Analysis Of Current Trends, Advances And Challenges Of Machine Learning (Ml) And Knowledge Extraction: From Ml To Explainable AI.* Industry Qualifications The Institute of Administrative Management UK, 2022. **58**: p. 54-62.
26. Tapeh, A.T.G. and M. Naser, *Artificial intelligence, machine learning, and deep learning in structural engineering: a scientometrics review of trends and best practices.* Archives of Computational Methods in Engineering, 2023. **30**(1): p. 115-159.
27. Heidari, A., N.J. Navimipour, and M. Unal, *Applications of ML/DL in the management of smart cities and societies based on new trends in information technologies: A systematic literature review.* Sustainable Cities and Society, 2022: p. 104089.
28. Sharma, M. and P. Garg, *Computational approaches for enzyme functional class prediction: a review.* Current Proteomics, 2014. **11**(1): p. 17-22.
29. Wang, Z., et al., *Review of protein subcellular localization prediction.* Current Bioinformatics, 2014. **9**(3): p. 331-342.
30. Murakami, Y. and K. Mizuguchi, *Applying the Naïve Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites.* Bioinformatics, 2010. **26**(15): p. 1841-1848.
31. Messih, M.A., et al., *Protein domain recurrence and order can enhance prediction of protein functions.* Bioinformatics, 2012. **28**(18): p. i444-i450.
32. Cozzetto, D., et al. *Protein function prediction by massive integration of evolutionary analyses and multiple data sources*. in *BMC bioinformatics*. 2013. Springer.
33. Mandle, A.K., P. Jain, and S.K. Shrivastava, *Protein structure prediction using support vector machine.* International Journal on Soft Computing, 2012. **3**(1): p. 67.

34. Liu, T., X. Zheng, and J. Wang, *Prediction of protein structural class for low-similarity sequences using support vector machine and PSI-BLAST profile.* Biochimie, 2010. **92**(10): p. 1330-1334.
35. Lise, S., et al., *Predictions of hot spot residues at protein-protein interfaces using support vector machines.* PLoS one, 2011. **6**(2): p. e16774.
36. Patel, M. and H. Shah. *Protein secondary structure prediction using support vector machines (svms)*. in *2013 International Conference on Machine Intelligence and Research Advancement*. 2013. IEEE.
37. Shen, Y. and A. Bax, *SPARTA+: a modest improvement in empirical NMR chemical shift prediction by means of an artificial neural network.* Journal of biomolecular NMR, 2010. **48**(1): p. 13-22.
38. Lan, L., et al. *MS-k NN: protein function prediction by integrating multiple data sources*. in *BMC bioinformatics*. 2013. BioMed Central.
39. Silla Jr, C.N. and A.A. Freitas, *Selecting different protein representations and classification algorithms in hierarchical protein function prediction.* Intelligent Data Analysis, 2011. **15**(6): p. 979-999.
40. Wallach, I., M. Dzamba, and A. Heifets, *AtomNet: a deep convolutional neural network for bioactivity prediction in structure-based drug discovery.* arXiv preprint arXiv:1510.02855, 2015.
41. Conduit, P.T., et al., *The centrosome-specific phosphorylation of Cnn by Polo/Plk1 drives Cnn scaffold assembly and centrosome maturation.* Developmental cell, 2014. **28**(6): p. 659-669.
42. Cao, R., et al., *ProLanGO: protein function prediction using neural machine translation based on a recurrent neural network.* Molecules, 2017. **22**(10): p. 1732.
43. Li, Z. and Y. Yu, *Protein secondary structure prediction using cascaded convolutional and recurrent neural networks.* arXiv preprint arXiv:1604.07176, 2016.
44. Sønderby, S.K., et al. *Convolutional LSTM networks for subcellular localization of proteins*. in *International conference on algorithms for computational biology*. 2015. Springer.
45. Liu, X., *Deep recurrent neural network for protein function prediction from sequence.* arXiv preprint arXiv:1701.08318, 2017.
46. Gligorijević, V., et al., *Structure-based protein function prediction using graph convolutional networks.* Nature communications, 2021. **12**(1): p. 1-14.
47. Lai, B. and J. Xu, *Accurate protein function prediction via graph attention networks with predicted structure information.* Briefings in Bioinformatics, 2022. **23**(1): p. bbab502.
48. Bonetta, R. and G. Valentino, *Machine learning techniques for protein function prediction.* Proteins: Structure, Function, and Bioinformatics, 2020. **88**(3): p. 397-413.
49. Mansoor, M., et al., *Gene Ontology GAN (GOGAN): a novel architecture for protein function prediction.* Soft Computing, 2022: p. 1-15.
50. Sharma, B., et al., *Accurate Clinical Toxicity Prediction using Multi-task Deep Neural Nets and Contrastive Molecular Explanations.* arXiv preprint arXiv:2204.06614, 2022.
51. Mills, C.L., P.J. Beuning, and M.J. Ondrechen, *Biochemical functional predictions for protein structures of unknown or uncertain function.* Computational and Structural Biotechnology Journal, 2015. **13**: p. 182-191.
52. Du, Z. and Y. Li, *Review and perspective on bioactive peptides: A roadmap for research, development, and future opportunities.* Journal of Agriculture and Food Research, 2022: p. 100353.
53. He, L., et al., *Fruit yield prediction and estimation in orchards: A state-of-the-art comprehensive review for both direct and indirect methods.* Computers and Electronics in Agriculture, 2022. **195**: p. 106812.
54. Jiang, Y., et al., *An expanded evaluation of protein function prediction methods shows an improvement in accuracy.* Genome biology, 2016. **17**(1): p. 1-19.
55. Zhou, N., et al., *The CAFA challenge reports improved protein function prediction and new functional annotations for hundreds of genes through experimental screens.* Genome biology, 2019. **20**(1): p. 1-23.
56. Peng, Z., et al., *Protein structure prediction in the deep learning era.* Current Opinion in Structural Biology, 2022. **77**: p. 102495.
57. de Vries, S.J. and A.M. Bonvin, *How proteins get in touch: interface prediction in the study of biomolecular complexes.* Current protein and peptide science, 2008. **9**(4): p. 394-406.
58. Shaheen, M., et al., *Applications of federated learning; Taxonomy, challenges, and research trends.* Electronics, 2022. **11**(4): p. 670.
59. Sharma, S. and P.K. Mandal, *A comprehensive report on machine learning-based early detection of alzheimer's disease using multi-modal neuroimaging data.* ACM Computing Surveys (CSUR), 2022. **55**(2): p. 1-44.
60. Luo, H., et al., *Combinations of Feature Selection and Machine Learning Algorithms for Object-Oriented Betel Palms and Mango Plantations Classification Based on Gaofen-2 Imagery.* Remote Sensing, 2022. **14**(7): p. 1757.
61. Yaprakdal, F. and B. Fatih, *Comparison of Robust Machine-learning and Deep-learning Models for Midterm Electrical Load Forecasting.* European Journal of Technique (EJT). **12**(2): p. 102-107.
62. Alshboul, O., et al., *Deep and machine learning approaches for forecasting the residual value of heavy construction equipment: A management decision support model.* Engineering, Construction and Architectural Management, 2021.
63. Ovek, D., et al., *Artificial intelligence based methods for hot spot prediction.* Current Opinion in Structural Biology, 2022. **72**: p. 209-218.
64. Quadrini, M., S. Daberdaku, and C. Ferrari, *Hierarchical representation for PPI sites prediction.* BMC bioinformatics, 2022. **23**(1): p. 1-34.
65. Brownlee, J., *Data preparation for machine learning*. 2022.
66. Ullah, I., et al., *A comparative performance of machine learning algorithm to predict electric vehicles energy consumption: A path towards sustainability.* Energy & Environment, 2022. **33**(8): p. 1583-1612.
67. Mishra, A., S.S. Mohapatra, and S.K. Bisoy, *Effective Deep Learning Algorithms for Personalized Healthcare Services*, in *Augmented Intelligence in Healthcare: A Pragmatic and Integrated Analysis*. 2022, Springer. p. 121-141.
68. Xiouras, C., et al., *Applications of artificial intelligence and machine learning algorithms to crystallization.* Chemical Reviews, 2022. **122**(15): p. 13006-13042.
69. Mou, M., et al., *Application of machine learning in spatial proteomics.* Journal of Chemical Information and Modeling, 2022. **62**(23): p. 5875-5895.
70. Lang, T., *The 'study walk-through': a method to help translate your protocol into an accurate and successful study.* 2022.

71. Brownlee, J., *Data Preparation for Machine Learning: Data Cleaning, Feature Selection, and Data Transforms in Python*. 2020: Machine Learning Mastery.
72. Bernardo, A. and E. Della Valle, *An extensive study of C-SMOTE, a Continuous Synthetic Minority Oversampling Technique for Evolving Data Streams.* Expert Systems with Applications, 2022. **196**: p. 116630.
73. Maulidevi, N.U. and K. Surendro, *SMOTE-LOF for noise identification in imbalanced data classification.* Journal of King Saud University-Computer and Information Sciences, 2022. **34**(6): p. 3413-3423.
74. Richhariya, B., et al., *Diagnosis of Alzheimer's disease using universum support vector machine based recursive feature elimination (USVM-RFE).* Biomedical Signal Processing and Control, 2020. **59**: p. 101903.
75. Kshirsagar, P., N. Balakrishnan, and A.D. Yadav, *Modelling of optimised neural network for classification and prediction of benchmark datasets.* Computer Methods in Biomechanics and Biomedical Engineering: Imaging & Visualization, 2020. **8**(4): p. 426-435.
76. Luengo, J., et al., *Big data preprocessing.* Cham: Springer, 2020.
77. Isanbaev, V., et al., *A comparative study on pretreatment methods and dimensionality reduction techniques for energy data disaggregation in home appliances.* Advanced Engineering Informatics, 2022. **54**: p. 101805.
78. Espadoto, M., et al., *Toward a quantitative survey of dimension reduction techniques.* IEEE transactions on visualization and computer graphics, 2019. **27**(3): p. 2153-2173.
79. Brownlee, J., *Statistical methods for machine learning: Discover how to transform data into knowledge with Python*. 2018: Machine Learning Mastery.
80. Janiesch, C., P. Zschech, and K. Heinrich, *Machine learning and deep learning.* Electronic Markets, 2021. **31**(3): p. 685-695.
81. Alam, S., et al., *One-class support vector classifiers: A survey.* Knowledge-Based Systems, 2020. **196**: p. 105754.
82. Faridoon, A., et al., *Combining SVM and ECOC for identification of protein complexes from protein protein interaction networks by integrating amino acids' physical properties and complex topology.* Interdisciplinary Sciences: Computational Life Sciences, 2020. **12**(3): p. 264-275.
83. Tucker, S., et al., *Distinct seasonal infectious agent profiles in life-history variants of juvenile Fraser River Chinook salmon: An application of high-throughput genomic screening.* PLoS One, 2018. **13**(4): p. e0195472.
84. Di Fiore, A., et al., *Human carbonic anhydrases and post-translational modifications: a hidden world possibly affecting protein properties and functions.* Journal of Enzyme Inhibition and Medicinal Chemistry, 2020. **35**(1): p. 1450-1461.
85. Rives, A., et al., *Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences.* Proceedings of the National Academy of Sciences, 2021. **118**(15): p. e2016239118.
86. Shatnawi, M., *Review of recent protein-protein interaction techniques.* Emerging Trends in Computational Biology, Bioinformatics, and Systems Biology: Algorithms and Software Tools, 2015.
87. Thangamurugan, S., M. Hollander, and V. Helms, *Identification of Putative Protein Complexes in Protein–Protein Interaction Networks.* Protein Interactions: The Molecular Basis of Interactomics, 2022: p. 77-99.
88. Zhang, F., et al., *DeepFunc: a deep learning framework for accurate prediction of protein functions from protein sequences and interactions.* Proteomics, 2019. **19**(12): p. 1900019.
89. Mousavian, Z., et al., *StrongestPath: a Cytoscape application for protein–protein interaction analysis.* BMC bioinformatics, 2021. **22**(1): p. 1-14.
90. Sun, X., et al., *RBPro-RF: use Chou's 5-steps rule to predict RNA-binding proteins via random forest with elastic net.* Chemometrics and Intelligent Laboratory Systems, 2020. **197**: p. 103919.
91. Jin, S., et al., *Application of deep learning methods in biological networks.* Briefings in bioinformatics, 2021. **22**(2): p. 1902-1917.
92. Marquet, C., et al., *Embeddings from protein language models predict conservation and variant effects.* Human genetics, 2022. **141**(10): p. 1629-1647.
93. Alsanousi, W.A., et al., *A novel deep learning-assisted hybrid network for plasmodium falciparum parasite mitochondrial proteins classification.* Plos one, 2022. **17**(10): p. e0275195.
94. Geçkın, D. and G.K. Demir. *Sequence Based Prediction of Protein-Protein Interactions via Siamese Neural Networks*. in *2022 Medical Technologies Congress (TIPTEKNO)*. 2022. IEEE.
95. Bhat, H.F. and M.A. Wani, *PSSM amino-acid composition based rules for gene identification.* International Journal of Advanced Technology and Engineering Exploration, 2018. **5**(46): p. 318-325.
96. Chauhan, S. and S. Ahmad, *Enabling full‐length evolutionary profiles based deep convolutional neural network for predicting DNA‐binding proteins from sequence.* Proteins: Structure, Function, and Bioinformatics, 2020. **88**(1): p. 15-30.
97. Kong, L. and L. Zhang, *An ensemble method for multi-type gram-negative bacterial secreted protein prediction by integrating different PSSM-based features.* SAR and QSAR in Environmental Research, 2019. **30**(3): p. 181-194.
98. Gonçalves, D.M., R. Henriques, and R.S. Costa, *Predicting postoperative complications in cancer patients: a survey bridging classical and machine learning contributions to postsurgical risk analysis.* Cancers, 2021. **13**(13): p. 3217.
99. Nápoles, G., et al., *Recurrence-aware long-term cognitive network for explainable pattern classification.* IEEE Transactions on Cybernetics, 2022.
100. Sidey-Gibbons, J.A. and C.J. Sidey-Gibbons, *Machine learning in medicine: a practical introduction.* BMC medical research methodology, 2019. **19**(1): p. 1-18.
101. Öztürk, H., et al., *Exploring chemical space using natural language processing methodologies for drug discovery.* Drug Discovery Today, 2020. **25**(4): p. 689-705.
102. Li, I., et al., *Neural natural language processing for unstructured data in electronic health records: A review.* Computer Science Review, 2022. **46**: p. 100511.
103. Li, F., et al., *Deep learning-based kcat prediction enables improved enzyme-constrained model reconstruction.* Nature Catalysis, 2022. **5**(8): p. 662-672.
104. You, R., X. Huang, and S. Zhu, *DeepText2GO: Improving large-scale protein function prediction with deep semantic text representation.* Methods, 2018. **145**: p. 82-90.
105. Dai, S., et al., *FullMeSH: improving large-scale MeSH indexing with full text.* Bioinformatics, 2020. **36**(5): p. 1533-1541.
106. Grigson, S.R., et al., *Organizing the bacterial annotation space with amino acid sequence embeddings.* BMC bioinformatics, 2022. **23**(1): p. 1-14.

107. Yin, R., et al., *A framework for predicting variable-length epitopes of human-adapted viruses using machine learning methods.* Briefings in Bioinformatics, 2022. **23**(5).
108. Rayan, R.A., I. Zafar, and C. Tsagkaris, *Artificial Intelligence and Big Data Solutions for COVID-19*, in *Intelligent Data Analysis for COVID-19 Pandemic*. 2021, Springer. p. 115-127.
109. Zafar, I., et al., *Genome-Wide Identification and Expression Analysis of PPOs and POX Gene Families in the Selected Plant Species.* Biosciences Biotechnology Research Asia, 2020. **17**(2): p. 301-318.
110. Zafar, I., et al., *Genome-wide identification and analysis of GRF (growth-regulating factor) gene family in Camila sativa through in silico approaches.* Journal of King Saud University-Science, 2022. **34**(4): p. 102038.
111. Ouyang, W., et al., *Analysis of the human protein atlas image classification competition.* Nature methods, 2019. **16**(12): p. 1254-1261.
112. Niu, L., et al., *Dynamic human liver proteome atlas reveals functional insights into disease pathways.* Molecular systems biology, 2022. **18**(5): p. e10947.
113. Rayan, R.A., et al., *Big data analytics for health: a comprehensive review of techniques and applications.* Big Data Analytics for Healthcare, 2022: p. 83-92.
114. Mao, W., J. He, and M.J. Zuo, *Predicting remaining useful life of rolling bearings based on deep feature representation and transfer learning.* IEEE Transactions on Instrumentation and Measurement, 2019. **69**(4): p. 1594-1608.
115. Wang, Y., et al., *Deep learning for fault-relevant feature extraction and fault classification with stacked supervised auto-encoder.* Journal of Process Control, 2020. **92**: p. 79-89.
116. Silva, A.B.O.V. and E.J. Spinosa, *Graph convolutional auto-encoders for predicting novel lncRNA-disease associations.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2021.
117. Abdelbaky, I., H. Tayara, and K.T. Chong, *Identification of miRNA-Small Molecule Associations by Continuous Feature Representation Using Auto-Encoders.* Pharmaceutics, 2021. **14**(1): p. 3.
118. Höglund, A., et al., *MultiLoc: prediction of protein subcellular localization using N-terminal targeting sequences, sequence motifs and amino acid composition.* Bioinformatics, 2006. **22**(10): p. 1158-1165.
119. Shatkay, H., et al., *SherLoc: high-accuracy prediction of protein subcellular localization by integrating text and protein sequence data.* Bioinformatics, 2007. **23**(11): p. 1410-1417.
120. You, Z.-H., et al. *Prediction of protein-protein interactions from amino acid sequences with ensemble extreme learning machines and principal component analysis*. in *BMC bioinformatics*. 2013. Springer.
121. Wang, J., et al., *Protein-protein interactions prediction using a novel local conjoint triad descriptor of amino acid sequences.* International journal of molecular sciences, 2017. **18**(11): p. 2373.
122. Dalkiran, A., et al., *ECPred: a tool for the prediction of the enzymatic functions of protein sequences based on the EC nomenclature.* BMC bioinformatics, 2018. **19**(1): p. 1-13.
123. Srivastava, A., A. Mahmood, and R. Srivastava. *A Comparative Analysis of SVM Random Forest Methods for Protein Function Prediction*. in *2017 International Conference on Current Trends in Computer, Electrical, Electronics and Communication (CTCEEC)*. 2017. IEEE.
124. Cai, C., et al., *SVM-Prot: web-based support vector machine software for functional classification of a protein from its primary sequence.* Nucleic acids research, 2003. **31**(13): p. 3692-3697.
125. Lee, B.J. and K.H. Ryu. *Feature extraction from protein sequences and classification of enzyme function*. in *2008 International Conference on BioMedical Engineering and Informatics*. 2008. IEEE.
126. Rahman, S.A., Z.A.M. Hussein, and A.A. Bakar. *Experimental study of different FSAs in classifying protein function*. in *2009 International Conference of Soft Computing and Pattern Recognition*. 2009. IEEE.
127. Li, F., et al., *GlycoMine: a machine learning-based approach for predicting N-, C-and O-linked glycosylation in the human proteome.* Bioinformatics, 2015. **31**(9): p. 1411-1419.
128. Acquaah-Mensah, G.K., S.M. Leach, and C. Guda, *Predicting the subcellular localization of human proteins using machine learning and exploratory data analysis.* Genomics, proteomics & bioinformatics, 2006. **4**(2): p. 120-133.
129. Kulmanov, M., M.A. Khan, and R. Hoehndorf, *DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier.* Bioinformatics, 2018. **34**(4): p. 660-668.
130. Li, Y.H., et al., *SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity.* PloS one, 2016. **11**(8): p. e0155290.
131. Nauman, M., et al., *Beyond homology transfer: Deep learning for automated annotation of proteins.* Journal of Grid Computing, 2019. **17**(2): p. 225-237.
132. Sun, T., et al., *Sequence-based prediction of protein protein interaction using a deep-learning algorithm.* BMC bioinformatics, 2017. **18**(1): p. 1-8.
133. Wang, Y.-B., et al., *Predicting protein–protein interactions from protein sequences by a stacked sparse autoencoder deep neural network.* Molecular BioSystems, 2017. **13**(7): p. 1336-1344.
134. You, R., et al., *GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank.* Bioinformatics, 2018. **34**(14): p. 2465-2473.
135. You, Z.-H., et al. *Prediction of protein-protein interactions from amino acid sequences using a novel multi-scale continuous and discontinuous feature set*. in *BMC bioinformatics*. 2014. Springer.
136. De Ferrari, L. and J.B. Mitchell, *From sequence to enzyme mechanism using multi-label machine learning.* BMC bioinformatics, 2014. **15**(1): p. 1-13.
137. Lou, W., et al., *Sequence based prediction of DNA-binding proteins based on hybrid feature selection using random forest and Gaussian naive Bayes.* PloS one, 2014. **9**(1): p. e86703.
138. Wang, H., et al., *From protein sequence to protein function via multi-label linear discriminant analysis.* IEEE/ACM transactions on computational biology and bioinformatics, 2016. **14**(3): p. 503-513.
139. Makrodimitris, S., R.C. van Ham, and M.J. Reinders, *Improving protein function prediction using protein sequence and GO-term similarities.* Bioinformatics, 2019. **35**(7): p. 1116-1124.
140. Resende, W.K., et al. *The use of support vector machine and genetic algorithms to predict protein function*. in *2012 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*. 2012. IEEE.
141. Wang, W., et al. *Protein function prediction based on physiochemical properties and protein granularity*. in *2013 IEEE International Conference on Granular Computing (GrC)*. 2013. IEEE.
142. Rahman, S.A., Z.A.M. Hussein, and A.A. Bakar. *Data mining framework for protein function prediction*. in *2008 International Symposium on Information Technology*. 2008. IEEE.

143. Silva, M.F.M., L.F. Leijoto, and C.N. Nobre, *Algorithms analysis in adjusting the SVM parameters: An approach in the prediction of protein function.* Applied Artificial Intelligence, 2017. **31**(4): p. 316-331.
144. Cai, C., et al., *Protein function classification via support vector machine approach.* Mathematical biosciences, 2003. **185**(2): p. 111-122.
145. Lee, B.J., et al. *Feature Extraction in Spatially-Conserved Regions and Protein Functional Classification*. in *2007 Frontiers in the Convergence of Bioscience and Information Technologies*. 2007. IEEE.
146. Zou, Z., et al., *mldeepre: Multi-functional enzyme function prediction with hierarchical multi-label deep learning.* Frontiers in genetics, 2019. **9**: p. 714.
147. Rice, S.B., G. Nenadic, and B.J. Stapley, *Mining protein function from text using term-based support vector machines.* BMC bioinformatics, 2005. **6**(1): p. 1-11.
148. Wong, A. and H. Shatkay. *Protein function prediction using text-based features extracted from the biomedical literature: the CAFA challenge*. in *BMC bioinformatics*. 2013. Springer.
149. Zheng, W. and C. Blake, *Using distant supervised learning to identify protein subcellular localizations from full-text scientific articles.* Journal of biomedical informatics, 2015. **57**: p. 134-144.
150. Funk, C.S., et al., *Evaluating a variety of text-mined features for automatic protein function prediction with GOstruct.* Journal of biomedical semantics, 2015. **6**(1): p. 1-14.
151. Shao, W., M. Liu, and D. Zhang, *Human cell structure-driven model construction for predicting protein subcellular location from biological images.* Bioinformatics, 2016. **32**(1): p. 114-121.
152. Gligorijević, V., M. Barot, and R. Bonneau, *deepNF: deep network fusion for protein function prediction.* Bioinformatics, 2018. **34**(22): p. 3873-3881.
153. Wei, L., et al., *Prediction of human protein subcellular localization using deep learning.* Journal of Parallel and Distributed Computing, 2018. **117**: p. 212-217.
154. Fa, R., et al., *Predicting human protein function with multi-task deep neural networks.* PloS one, 2018. **13**(6): p. e0198216.
155. Wang, Y. and T. Li, *Local feature selection based on artificial immune system for classification.* Applied Soft Computing, 2020. **87**: p. 105989.
156. Kim, W., et al., *Electricity load forecasting using advanced feature selection and optimal deep learning model for the variable refrigerant flow systems.* Energy Reports, 2020. **6**: p. 2604-2618.
157. Khaire, U.M. and R. Dhanalakshmi, *Stability of feature selection algorithm: A review.* Journal of King Saud University-Computer and Information Sciences, 2022. **34**(4): p. 1060-1073.
158. Brownlee, J., *How to choose a feature selection method for machine learning.* Machine Learning Mastery, 2019. **10**.
159. Molina, L.C., L. Belanche, and À. Nebot. *Feature selection algorithms: A survey and experimental evaluation*. in *2002 IEEE International Conference on Data Mining, 2002. Proceedings.* 2002. IEEE.
160. Saeys, Y., I. Inza, and P. Larranaga, *A review of feature selection techniques in bioinformatics.* bioinformatics, 2007. **23**(19): p. 2507-2517.
161. Wang, L., Y. Wang, and Q. Chang, *Feature selection methods for big data bioinformatics: A survey from the search perspective.* Methods, 2016. **111**: p. 21-31.
162. Sapoval, N., et al., *Current progress and open challenges for applying deep learning across the biosciences.* Nature Communications, 2022. **13**(1): p. 1728.
163. Greener, J.G., et al., *A guide to machine learning for biologists.* Nature Reviews Molecular Cell Biology, 2022. **23**(1): p. 40-55.
164. Görmez, Y. and Z. Aydin, *IGPRED-MultiTask: a deep learning model to predict protein secondary structure, torsion angles and solvent accessibility*. IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2022.
165. Lim, H., et al., *Artificial intelligence approaches to human-microbiome protein–protein interactions.* Current Opinion in Structural Biology, 2022. **73**: p. 102328.
166. Valavi, R., et al., *Predictive performance of presence‐only species distribution models: a benchmark study with reproducible code.* Ecological Monographs, 2022. **92**(1): p. e01486.
167. Wu, R., et al., *High-resolution de novo structure prediction from primary sequence*. BioRxiv, 2022: p. 2022.07. 21.500999.
168. Antonelli, M., et al., *The medical segmentation decathlon.* Nature communications, 2022. **13**(1): p. 4128.
169. Koumi, F., M. Aldasht, and H. Tamimi. *Efficient feature selection using particle swarm optimization: a hybrid filters-wrapper approach*. in *2019 10th international conference on information and communication systems (ICICS)*. 2019. IEEE.
170. Ghosh, M., et al., *A wrapper-filter feature selection technique based on ant colony optimization.* Neural Computing and Applications, 2020. **32**(12): p. 7839-7857.
171. Fernandez-Lozano, C., et al., *Classification of signaling proteins based on molecular star graph descriptors using Machine Learning models.* Journal of theoretical biology, 2015. **384**: p. 50-58.
172. Yang, R., et al., *A novel feature extraction method with feature selection to identify Golgi-resident protein types from imbalanced data.* International journal of molecular sciences, 2016. **17**(2): p. 218.
173. Lin, H., et al., *Prediction of subcellular location of mycobacterial protein using feature selection techniques.* Molecular diversity, 2010. **14**(4): p. 667-671.
174. Accorsi, R., et al., *Data mining and machine learning for condition-based maintenance.* Procedia Manufacturing, 2017. **11**: p. 1153-1161.
175. Tang, H., et al., *HBPred: a tool to identify growth hormone-binding proteins.* International journal of biological sciences, 2018. **14**(8): p. 957.
176. Al-Shahib, A., R. Breitling, and D. Gilbert, *FrankSum: new feature selection method for protein function prediction.* International journal of neural systems, 2005. **15**(04): p. 259-275.
177. Rahman, S.A., A.A. Bakar, and Z.A.M. Hussein. *Feature selection and classification of protein subfamilies using rough sets*. in *2009 International Conference on Electrical Engineering and Informatics*. 2009. IEEE.
178. Ding, C. and H. Peng, *Minimum redundancy feature selection from microarray gene expression data.* Journal of bioinformatics and computational biology, 2005. **3**(02): p. 185-205.
179. Hu, L., et al., *Predicting functions of proteins in mouse based on weighted protein-protein interaction network and protein hybrid properties.* PloS one, 2011. **6**(1): p. e14556.

180. Kumar, C., G. Li, and A. Choudhary. *Enzyme function classification using protein sequence features and random forest*. in *2009 3rd International Conference on Bioinformatics and Biomedical Engineering*. 2009. IEEE.
181. Clark, W.T. and P. Radivojac, *Analysis of protein function and its prediction from amino acid sequence.* Proteins: Structure, Function, and Bioinformatics, 2011. **79**(7): p. 2086-2096.
182. Dos Santos, B.C., C.N. Nobre, and L.E. Zárate. *Multi-objective genetic algorithm for feature selection in a protein function prediction context*. in *2018 IEEE Congress on Evolutionary Computation (CEC)*. 2018. IEEE.
183. Yao, L., et al., *Detection of coronavirus in environmental surveillance and risk monitoring for pandemic control.* Chemical Society Reviews, 2021. **50**(6): p. 3656-3676.
184. Kulmanov, M. and R. Hoehndorf, *DeepGOPlus: improved protein function prediction from sequence.* Bioinformatics, 2020. **36**(2): p. 422-429.
185. You, R., et al., *NetGO: improving large-scale protein function prediction with massive network information.* Nucleic acids research, 2019. **47**(W1): p. W379-W387.
186. Rifaioglu, A.S., et al., *DEEPred: automated protein function prediction with multi-task feed-forward deep neural networks.* Scientific reports, 2019. **9**(1): p. 1-16.
187. Cai, Y., J. Wang, and L. Deng, *SDN2GO: An Integrated Deep Learning Model for Protein Function Prediction.* Frontiers in bioengineering and biotechnology, 2020. **8**: p. 391.
188. Zacharaki, E.I., *Prediction of protein function using a deep convolutional neural network ensemble.* PeerJ Computer Science, 2017. **3**: p. e124.
189. Gligorijevic, V., et al., *Structure-based function prediction using graph convolutional networks.* bioRxiv, 2020: p. 786236.
190. Du, Z., et al., *DeepAdd: Protein function prediction from k-mer embedding and additional features.* Computational Biology and Chemistry, 2020. **89**: p. 107379.
191. Miranda, L.J. and J. Hu. *A Deep Learning Approach Based on Stacked Denoising Autoencoders for Protein Function Prediction*. in *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*. 2018. IEEE.
192. Giri, S.J., et al., *MultiPredGO: Deep Multi-Modal Protein Function Prediction by Amalgamating Protein Structure, Sequence, and Interaction.* IEEE Journal of Biomedical and Health Informatics, 2020.
193. Consortium, G.O., *Gene ontology consortium: going forward.* Nucleic acids research, 2015. **43**(D1): p. D1049-D1056.
194. Pruitt, K.D., T. Tatusova, and D.R. Maglott, *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.* Nucleic acids research, 2007. **35**(suppl_1): p. D61-D65.
195. Consortium, U., *UniProt: a hub for protein information.* Nucleic Acids Res, 2015. **43**(D1): p. D204-D212.
196. Berman, H., K. Henrick, and H. Nakamura, *Announcing the worldwide protein data bank.* Nature Structural & Molecular Biology, 2003. **10**(12): p. 980-980.
197. Velankar, S., et al., *PDBe: improved accessibility of macromolecular structure data from PDB and EMDB.* Nucleic acids research, 2016. **44**(D1): p. D385-D395.
198. Berman, H.M., et al., *The protein data bank.* Nucleic acids research, 2000. **28**(1): p. 235-242.
199. Ulrich, E., et al., *BioMagResBank Nucleic Acids Res. 36.* D402–D408, 2008.
200. Liu, L., et al., *Predicting protein function via multi-label supervised topic model on gene ontology.* Biotechnology & Biotechnological Equipment, 2017. **31**(3): p. 630-638.
201. Cheng, X., et al., *pLoc_bal-mAnimal: predict subcellular localization of animal proteins by balancing training dataset and PseAAC.* Bioinformatics, 2019. **35**(3): p. 398-406.
202. Ni, Q., et al. *Using logistic regression method to predict protein function from protein-protein interaction data*. in *2009 3rd International Conference on Bioinformatics and Biomedical Engineering*. 2009. IEEE.
203. Maheshwari, S. and M. Brylinski, *Prediction of protein–protein interaction sites from weakly homologous template structures using meta‐threading and machine learning.* Journal of Molecular Recognition, 2015. **28**(1): p. 35-48.
204. de Campos Merschmann, L.H. and A.A. Freitas. *An extended local hierarchical classifier for prediction of protein and gene functions*. in *International Conference on Data Warehousing and Knowledge Discovery*. 2013. Springer.
205. Fabris, F. and A.A. Freitas. *An Efficient Algorithm for Hierarchical Classification of Protein and Gene Functions*. in *2014 25th International Workshop on Database and Expert Systems Applications*. 2014. IEEE.
206. Yang, X., et al., *Concepts of Artificial Intelligence for Computer-Assisted Drug Discovery.* Chemical Reviews, 2019. **119**.
207. Hayat, M., A. Khan, and M. Yeasin, *Prediction of membrane proteins using split amino acid and ensemble classification.* Amino acids, 2012. **42**(6): p. 2447-2460.
208. Khan, N., et al., *Batteries state of health estimation via efficient neural networks with multiple channel charging profiles.* IEEE Access, 2020. **9**: p. 7797-7813.
209. Yu, C.S., C.J. Lin, and J.K. Hwang, *Predicting subcellular localization of proteins for Gram‐negative bacteria by support vector machines based on n‐peptide compositions.* Protein science, 2004. **13**(5): p. 1402-1406.
210. Park, K.-J. and M. Kanehisa, *Prediction of protein subcellular locations by support vector machines using compositions of amino acids and amino acid pairs.* Bioinformatics, 2003. **19**(13): p. 1656-1663.
211. Zhou, X.-B., et al., *Using Chou's amphiphilic pseudo-amino acid composition and support vector machine for prediction of enzyme subfamily classes.* Journal of theoretical biology, 2007. **248**(3): p. 546-551.
212. Cai, Y.-D., et al., *Support vector machines for predicting protein structural class.* BMC bioinformatics, 2001. **2**(1): p. 1-5.
213. Lanckriet, G.R., et al., *Kernel-based data fusion and its application to protein function prediction in yeast*, in *Biocomputing 2004*. 2003, World Scientific. p. 300-311.
214. Cozzetto, D., et al., *FFPred 3: feature-based function prediction for all Gene Ontology domains.* Scientific reports, 2016. **6**(1): p. 1-11.
215. Tang, H., et al., *Identification of apolipoprotein using feature selection technique.* Scientific reports, 2016. **6**(1): p. 1-6.
216. Zhang, S.-B. and Q.-R. Tang, *Predicting protein subcellular localization based on information content of gene ontology terms.* Computational biology and chemistry, 2016. **65**: p. 1-7.
217. Badal, V.D., P.J. Kundrotas, and I.A. Vakser, *Natural language processing in text mining for structural modeling of protein complexes.* BMC bioinformatics, 2018. **19**(1): p. 1-10.
218. Breiman, L., *Random forests.* Machine learning, 2001. **45**(1): p. 5-32.

219. Peled, S., et al., *De-novo protein function prediction using DNA binding and RNA binding proteins as a test case.* Nature communications, 2016. **7**(1): p. 1-9.
220. Singh, M., P. Singh, and H. Singh. *Decision tree classifier for human protein function prediction*. in *2006 International Conference on Advanced Computing and Communications*. 2006. IEEE.
221. Vens, C., et al., *Decision trees for hierarchical multi-label classification.* Machine learning, 2008. **73**(2): p. 185.
222. Cerri, R., et al. *Multi-label Feature Selection Techniques for Hierarchical Multi-label Protein Function Prediction*. in *2018 International Joint Conference on Neural Networks (IJCNN)*. 2018. IEEE.
223. Sajjad, M., et al., *Towards efficient building designing: Heating and cooling load prediction via multi-output model.* Sensors, 2020. **20**(22): p. 6419.
224. Alsanousi, W.A., et al., *Review about off-line handwriting Arabic text recognition.* Int. J. Comput. Sci. Mob. Comput, 2017. **6**: p. 4-14.
225. Cerri, R., et al., *Reduction strategies for hierarchical multi-label classification in protein function prediction.* BMC bioinformatics, 2016. **17**(1): p. 1-24.
226. Nievola, J.C., E.C. Paraiso, and A.A. Freitas. *A hierarchical neural network for predicting protein functions*. in *2015 IEEE 15th International Conference on Bioinformatics and Bioengineering (BIBE)*. 2015. IEEE.
227. Smale, S., et al., *Mathematics of the neural response.* Foundations of Computational Mathematics, 2010. **10**(1): p. 67-91.
228. Yalamanchili, H.K., J. Wang, and Q.-W. Xiao. *NRProF: Neural response based protein function prediction algorithm*. in *2011 IEEE International Conference on Systems Biology (ISB)*. 2011. IEEE.
229. Mercado-Díaz, L.R., J. Navarro-García, and J.A. Jaramillo-Garzón. *A comparison of class-balance strategies for SVM in the problem of protein function prediction*. in *2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA)*. 2015. IEEE.
230. Ahmed, I., P. Witbooi, and A. Christoffels, *Prediction of human-Bacillus anthracis protein–protein interactions using multi-layer neural network.* Bioinformatics, 2018. **34**(24): p. 4159-4164.
231. Kerepesi, C., et al., *Prediction and characterization of human ageing-related proteins by using machine learning.* Scientific reports, 2018. **8**(1): p. 1-13.
232. AlQuraishi, M., *Machine learning in protein structure prediction.* Current opinion in chemical biology, 2021. **65**: p. 1-8.
233. Yu, T., et al., *Enzyme function prediction using contrastive learning.* Science, 2023. **379**(6639): p. 1358-1363.
234. Ullah, F.U.M., et al., *Diving deep into short-term electricity load forecasting: comparative analysis and a novel framework.* Mathematics, 2021. **9**(6): p. 611.
235. Friedman, J.H., *Greedy function approximation: a gradient boosting machine.* Annals of statistics, 2001: p. 1189-1232.
236. Freund, Y. and R.E. Schapire, *A decision-theoretic generalization of on-line learning and an application to boosting.* Journal of computer and system sciences, 1997. **55**(1): p. 119-139.
237. Khan, N., et al., *SD-Net: Understanding overcrowded scenes in real-time via an efficient dilated convolutional neural network.* Journal of Real-Time Image Processing, 2020: p. 1-15.
238. LeCun, Y., Y. Bengio, and G. Hinton, *Deep learning.* nature, 2015. **521**(7553): p. 436-444.
239. Khan, N., et al., *DB-Net: A novel dilated CNN based multi-step forecasting model for power consumption in integrated local energy systems.* International Journal of Electrical Power & Energy Systems, 2021: p. 107023.
240. Simonyan, K. and A. Zisserman, *Very deep convolutional networks for large-scale image recognition.* arXiv preprint arXiv:1409.1556, 2014.
241. Yang, Z., et al. *Combining partial parameter transfer strategy and spatial pyramid pooling for biological-cell classification*. in *Tenth International Conference on Graphics and Image Processing (ICGIP 2018)*. 2019. International Society for Optics and Photonics.
242. Lu, S., et al., *Efficient ResNet Model to Predict Protein-Protein Interactions With GPU Computing.* IEEE Access, 2020. **8**: p. 127834-127844.
243. Li, Z., et al., *Protein contact map prediction based on ResNet and DenseNet.* BioMed research international, 2020. **2020**.
244. Song, J., et al., *A Novel Prediction Method for ATP-Binding Sites From Protein Primary Sequences Based on Fusion of Deep Convolutional Neural Network and Ensemble Learning.* IEEE Access, 2020. **8**: p. 21485-21495.
245. Guo, Z., J. Hou, and J. Cheng, *DNSS2: improved ab initio protein secondary structure prediction using advanced deep learning architectures.* Proteins: Structure, Function, and Bioinformatics, 2021. **89**(2): p. 207-217.
246. Masurkar, S.R. and P.P. Rege. *Human Protein Subcellular Localization using Convolutional Neural Network as Feature Extractor*. in *2019 10th International Conference on Computing, Communication and Networking Technologies (ICCCNT)*. 2019. IEEE.
247. Krizhevsky, A., I. Sutskever, and G.E. Hinton, *ImageNet classification with deep convolutional neural networks.* Communications of the ACM, 2017. **60**(6): p. 84-90.
248. Taju, S.W., et al., *DeepEfflux: a 2D convolutional neural network model for identifying families of efflux proteins in transporters.* Bioinformatics, 2018. **34**(18): p. 3111-3117.
249. Quang, D. and X. Xie, *DanQ: a hybrid convolutional and recurrent deep neural network for quantifying the function of DNA sequences.* Nucleic acids research, 2016. **44**(11): p. e107-e107.
250. Ahmet, S.R., et al., *DEEPred: Automated Protein Function Prediction with Multi-task Feed-forward Deep Neural Networks.* Scientific Reports (Nature Publisher Group), 2019. **9**(1).
251. Wu, Z., et al., *CFAGO: cross-fusion of network and attributes based on attention mechanism for protein function prediction.* Bioinformatics, 2023. **39**(3): p. btad123.
252. Sanderson, T., et al., *ProteInfer, deep neural networks for protein functional inference.* Elife, 2023. **12**: p. e80942.
253. Haq, I.U., et al., *Sequential learning-based energy consumption prediction model for residential and commercial sectors.* Mathematics, 2021. **9**(6): p. 605.
254. Alsanousi, W.A., et al., *A novel deep learning-assisted hybrid network for plasmodium falciparum parasite mitochondrial proteins classification.* Plos one, 2022. **17**(10): p. e0275195.
255. Pearlmutter, B.A., *Learning state space trajectories in recurrent neural networks.* Neural Computation, 1989. **1**(2): p. 263-269.
256. Hou, J., *New approaches of protein function prediction from protein interaction networks*. 2017: Academic Press.

257. Amidi, S., et al., *Automatic single-and multi-label enzymatic function prediction by machine learning.* PeerJ, 2017. **5**: p. e3095.
258. Wu, J.-S., S.-J. Huang, and Z.-H. Zhou, *Genome-wide protein function prediction through multi-instance multi-label learning.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2014. **11**(5): p. 891-902.
259. Dietterich, T.G., R.H. Lathrop, and T. Lozano-Pérez, *Solving the multiple instance problem with axis-parallel rectangles.* Artificial intelligence, 1997. **89**(1-2): p. 31-71.
260. Sun, P., et al., *Protein function prediction using function associations in protein–protein interaction network.* IEEE Access, 2018. **6**: p. 30892-30902.
261. De Santis, E., et al. *Dissimilarity space representations and automatic feature selection for protein function prediction*. in *2018 International joint conference on neural networks (IJCNN)*. 2018. IEEE.
262. Ben-Hur, A. and D. Brutlag, *Sequence motifs: highly predictive features of protein function*, in *Feature extraction*. 2006, Springer. p. 625-645.
263. Ruepp, A., et al., *The FunCat, a functional annotation scheme for systematic classification of proteins from whole genomes.* Nucleic acids research, 2004. **32**(18): p. 5539-5545.
264. Webb, E., *International Union of Biochemistry and Molecular Biology, Nomenclature Committee. Enzyme Nomenclature 1992: Recommendations of the Nomenclature Committee of the International Union Of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. 1992, San Diego: Academic Press.
265. Yu, B., et al., *Accurate prediction of subcellular location of apoptosis proteins combining Chou's PseAAC and PsePSSM based on wavelet denoising.* Oncotarget, 2017. **8**(64): p. 107640.
266. Lu, Z., et al., *Predicting subcellular localization of proteins using machine-learned classifiers.* Bioinformatics, 2004. **20**(4): p. 547-556.
267. Li, Y., et al. *Prediction of Calmodulin-Binding Proteins Using Short-Linear Motifs*. in *International Conference on Bioinformatics and Biomedical Engineering*. 2017. Springer.
268. Fodeh, S., A. Tiwari, and H. Yu. *Exploiting PubMed for protein molecular function prediction via NMF based multi-label classification*. in *2017 IEEE International Conference on Data Mining Workshops (ICDMW)*. 2017. IEEE.
269. Lee, B.J., et al., *Identification of protein functions using a machine-learning approach based on sequence-derived properties.* Proteome science, 2009. **7**(1): p. 1-19.
270. Youngs, N., et al., *Negative example selection for protein function prediction: the NoGO database.* PLoS Comput Biol, 2014. **10**(6): p. e1003644.
271. Yu, G., et al., *Protein function prediction with incomplete annotations.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2013. **11**(3): p. 579-591.
272. Wu, J., et al. *Predicting protein functions of bacteria genomes via multi-instance multi-label active learning*. in *2018 IEEE 3rd International Conference on Integrated Circuits and Microsystems (ICICM)*. 2018. IEEE.
273. Guo, X., et al., *Human protein subcellular localization with integrated source and multi-label ensemble classifier.* Scientific Reports, 2016. **6**(1): p. 1-11.
274. Jiang, J.Q. and L.J. McQuay, *Predicting protein function by multi-label correlated semi-supervised learning.* IEEE/ACM Transactions on Computational Biology and Bioinformatics, 2011. **9**(4): p. 1059-1069.

