Identification Of Hormone Binding Proteins Using Multi-Informative Features Incorporating Ensemble Learning Approach

KHADIZA ISLAM PUNNO

Bachelor of Science

DAFFODIL INTERNATIONAL UNIVERSITY

DAFFODIL INTERNATIONAL UNIVERSITY

DECLARATION OF THESIS AND COPYRIGHT

Author's Full Name : KHADIZA ISLAM PUNNO

Date of Birth :

Title : Identification Of Hormone Binding Proteins Using Multi-Informative Features Incorporating Ensemble Learning Approach

Academic Session :

I declare that this thesis is classified as:

☐	CONFIDENTIAL	(Contains confidential information under the Official Secret Act 1997)*
☐	RESTRICTED	(Contains restricted information as specified by the organization where research was done)*
☐	OPEN ACCESS	I agree that my thesis to be published as online open access (Full Text)

I acknowledge that Daffodil International University reserves the following rights:

1. The Thesis is the Property of Daffodil International University.
2. The Library of Daffodil International University has the right to make copies of the thesis for the purpose of research only.
3. The Library of Daffodil International University has the right to make copies of the thesis for academic exchange.

Certified by:

_____ (Student's Signature)

_____ (Supervisor's Signature)

NOTE : * If the thesis is CONFIDENTIAL or RESTRICTED, please attach a thesis declaration letter.

THESIS DECLARATION LETTER (OPTIONAL)

Librarian,
Daffodil International University,
Daffodil Smart City,
Ashulia.Dhaka,Bangladesh

Dear Sir,

CLASSIFICATION OF THESIS AS RESTRICTED

Please be informed that the following thesis is classified as RESTRICTED for a period of three (3) years from the date of this letter. The reasons for this classification are as listed below.
Author's Name
Thesis Title

Reasons (i)

(ii)

(iii)

Thank you.

Yours faithfully,

_____
(Supervisor's Signature)

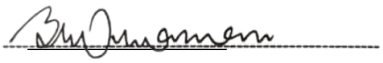
Date:

Stamp:

Note: This letter should be written by the supervisor and addressed to the Librarian, *Daffodil International University* with its copy attached to the thesis.

SUPERVISOR'S DECLARATION

I hereby declare that I have checked this thesis and, in my opinion, this thesis is adequate in terms of scope and quality for the award of the degree of Bachelor of Science.



(Supervisor's Signature)

Full Name : Mr. Khalid Been Badruzzaman Biplob

Position : Lecturer (Senior Scale)

Date : 14 November 2025

STUDENT'S DECLARATION

I hereby declare that the work in this thesis is based on my original work except for quotations and citations which have been duly acknowledged. I also declare that it has not been previously or concurrently submitted for any other degree at Daffodil International University or any other institution.

Khadiza Islam Punno

(Student's Signature)

Full Name : Khadiza Islam Punno

ID Number : 221-35-960

Date : 14 November 2025

Identification Of Hormone Binding Proteins Using Multi-Informative Features Incorporating Ensemble Learning Approach

KHADIZA ISLAM PUNNO

Thesis submitted in fulfillment of the requirements
for the award of the degree of
Bachelor of Science

Department of Software Engineering (Non-Major)

DAFFODIL INTERNATIONAL UNIVERSITY

DECEMBER 2025

APPROVAL

This thesis titled on "Identification of Hormone Binding Proteins Using Multi-Informative Features Incorporating Ensemble Learning Approach", submitted by **Student Name (ID: 221-35-960)** to the Department of Software Engineering, Daffodil International University has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of Bachelor of Science in Software Engineering and approval as to its style and contents.

BOARD OF EXAMINERS

Chairman

Dr. S. M. Hasan Mahmud
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 1

A.H.M Shahariar Parvez
Associate Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 2

Tapushe Rabaya Toma
Assistant Professor
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

Internal Examiner 3

Khalid Been md. Badruzzaman Biplob
Lecturer (Senior Scale)
Department of Software Engineering
Faculty of Science and Information Technology
Daffodil International University

External Examiner

Dr. Md Sazzadur Rahman
Professor
Institute of Information technology
Jahangirnagar University, Bangladesh

ACKNOWLEDGEMENTS

This thesis would not have been possible without the guidance, support, and patience of several people who have helped me throughout my academic journey. First and foremost, I would like to express my deepest gratitude to my thesis supervisor, **Khalid Been Md .Badruzzaman Biplop.** Your invaluable expertise, critical feedback, and constant encouragement were instrumental in shaping this research. Your insightful questions pushed me to think deeper and refine my methodology, and your open door provided unwavering support. Finally, I could not have completed this journey without the unconditional love and support of my family. To my parents, thank you for your endless belief in me, for your patience, and for your constant encouragement, especially when the challenges seemed overwhelming. This accomplishment is as much yours, as it is mine.

DEDICATION

Dedicated to my parents.

ABSTRACT

The growth of genomic and proteomic databases has led to an enormous annotation gap in which the amount of uncharacterised protein sequences has dramatically increased relative to the ability of slow and costly experimental classification processes. This will require creation of speedy, precise as well as scalable computing instruments. This thesis solves this challenge by undertaking an in-depth comparative analysis of the case to determine the best machine learning model to classify proteins using sequence-based derived features. An equal set of 3,570 protein sequences (1,785 in each class) was prefiltered and fed into a large-scale feature engineering pipeline with the ifeature library. A total of 15 different feature sets such as Amino Acid Composition (AAC), Dipeptide Composition (DPC), Tripeptide Composition (TPC) and a host of autocorrelation and physicochemical properties were concatenated to yield 11,466 dimensions each protein. Nine different machine learning and deep learning models were trained and heavily tested on this high-dimensional data: Support Vector Machine (SVM), K-Nearest Neighbors (KNN), Random Forest (RF), XGBoost (XGB), LightGBM (LGBM), a custom Ensemble (RF+XGB+DT), Artificial Neural Network (ANN), Recurrent Neural Network (RNN) and Convolutional Neural Network (CNN). The models were tested on a held-out test set. The findings showed that the random forest (RF) model was far the best classifier with high performance according to all the essential measures, as well as an Accuracy of 0.8163, Sensitivity of 0.8400, and an AUC of 0.8350. It is important to note that more complex boosting (XGB, LGBM) and deep learning (ANN, CNN) models performed worse, whereas the RNN architecture was unable to identify the meaningful patterns in the case of the fixed feature vector. This paper gives a concise, empirical reference point of this high dimensional bioinformatics exercise, and the determination was that the Random Forest classifier is the strongest and the most efficient model to this particular feature-based technique. This observation is a useful, practical suggestion to researchers creating such protein classification pipelines.

Keyword: Protein Sequence, Machine Learning, Deep Learning, Random Forest

TABLEOF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1. Background and Motivation

Proteins form the main laborers of life. These are complicated macromolecules built out of amino acid chains, which do almost all work of a biological cell. They are enzymes, catalyze essential chemical reactions; they constitute parts of the structure, providing body and shape to cells and tissues; they serve as signaling molecules, transmitting signals between cells; and they serve as defense, as antibodies. The specific role of a protein is tightly connected with its specific three-dimensional structure which is, in its turn, determined by its one-dimensional sequence of amino acids. It is one of the most basic problems of biology today to understand this sequence-structure-function paradigm.

Within recent decades, the development of genomic and proteomic technologies has presupposed a never-before-seen boom of biological data. Protein sequence information realized on a large scale by large-scale sequencing projects is much larger than our capacity to characterize any protein experimentally. This flood of data is a huge prospect, and is also a huge problem: how do we give functions or categories to these millions of hypothetical or uncharacterized protein sequences?

2. Problem Statement

It is unfeasibly slow, costly and resource-intensive to manually establish the role and type of each newly identified protein sequence using laboratory work (in vivo or in vitro). Such an annotationgap requires the creationof scalable, quick, and exact computationalapproaches.

Although older bioinformatics techniques such as sequence alignment (e.g., BLAST) have been essential, these techniques mostly depend on finding similarity in sequences. Such techniques fail to succeed in the classification of proteins with a shared purpose or family but separated drastically at the sequence level (low sequence homology). Thus, it is urgently necessary to develop more advanced computational methods that would be able to identify some deeper and more hidden patterns in the sequences to determine the class of a protein.

3. Significance and Expected Outcome

This project will be important as it will help in the analysis of high throughput of biological data. This research is expected to deliver a useful machine learning model to classify protein sequences quickly and cheaply through the creation of a powerful model.

It is anticipated that the result of this study will be a validated computational model, trained on the protein dataset.csv, which will be able to predict the protein classes with high accuracy. The downstreambenefits of such a model may include several benefits such as:

- **Accelerating Research:** Allowing researchers to quickly gain insights into the potential functions of newly sequenced proteins.
- **Informing Experimental Design:** Helping to prioritize proteins for further, in-depth experimental validation.
- **Biomedical Applications:** Contributing to the larger goal of understanding disease mechanisms, identifying potentialdrug targets, and designing novel therapeutics.

4. Aims and Objectives

The thesis will help solve this issue by designing a machine learning model to classify protein sequences. The basic idea of the hypothesis is that given raw amino acid sequences, a machine learning algorithm can be trained to predict the correct classification of a protein into various protein classes by transforming the information into a complete set of numerical features, which represent both their compositionand their physicochemical properties.

To achieve this aim, the following objectives have been set:

1. **To review** the existing literature on computational protein classification, focusing on machine learning techniques and sequence-based feature extraction methods.
2. **To preprocess** the protein_dataset.csv, ensuring it is clean and suitable for feature extraction and model training.
3. **To extract** a diverse set of sequence-based features (e.g., amino acid composition, dipeptide composition, physicochemical properties) using established bioinformatics libraries such as propy3 and ifeature.
4. **To train** a machine learning classifier (as identified in the notebook) using the extracted feature vectors.
5. **To rigorously evaluate** the trained model's predictive performance on an independent test set using a comprehensive suite of statistical metrics, including Accuracy, Sensitivity, Specificity, F1-Score, Matthews Correlation Coefficient (MCC), and the Area Under the ROC Curve (AUC).
6. **To analyze** the results, interpret the model's performance via its confusion matrix, and discuss its strengths and weaknesses.

1.5 Thesis Structure

Five chapters are used to organize this thesis. Chapter 2 presents a literature and background review of the concerned background and literature, including protein basics, classification systems, and machine learning in this field. Chapter 3 explains the entire methodology which includes the dataset, exact methods of feature extraction, architecture of the machine learning model and metrics of evaluation applied. Chapter 4 gives the result of the computational experiments, and then details the performance of the model are discussed and analyzed. Lastly, Chapter 5 serves as the conclusion of the thesis as it summarizes the main results and gives indications on where future research can be conducted.

CHAPTER 2

LITERATURE REVIEW

1. Introduction

This chapter includes the background knowledge and the setting of the research in this thesis. It gives a short introduction to the structure and classification of proteins, moves on to the known computational tools employed in analysis of proteins and finally the machine learning and feature extractiontechniques that are the main focus of this paper.

2. Protein Basics

Proteins are complex and large molecules which consist of smaller molecules known as amino acids. There are 20 standard amino acids which are united into chains, in a similar way as the beads on a string. The combination of these amino acids in certain order constitutes the inherent order of the protein, the primary structure.

This main chain is however not a bare line of sequence but takes complex, three-dimensional shapes. The secondary structure is made up of local folding patterns, including alpha-helices and beta-sheets. The tertiary structure of an individual amino acid chain is its 3D shape, which is critical in its functionality. Other proteins are composed of more than one chain (subunits) which form assemblies and this is what is referred to as the quaternary structure. The complicated nature of this relationship in which the 1D sequence governs the 3D structure, which governs the biological activity is a key belief in molecular biology.

3. Protein Classification

Given the vast diversity of proteins, classification is essential for organizing and understanding their roles. Proteins can be classified based on several criteria:

- **Function:** Grouping proteins by their biological role (e.g., enzymes, structural proteins, transport proteins, signaling proteins).
- **Family and Superfamily:** Grouping proteins based on evolutionary relatedness, inferred from similarities in their sequence and structure. Proteins in the same family often share a common function.
- **Subcellular Localization:** Classifying proteins based on where they reside and function within the cell (e.g., nucleus, cytoplasm, membrane).

Major biological databases serve as repositories for this information, with UniProt being a comprehensive resourcefor proteinsequence and functional annotation. Other databases like Pfam

(Protein families) and SCOP (Structural Classification of Proteins) provide specialized classification schemes.

4. The Shift from Experimental to Computational Methods

As stated in chapter one, the pace at which protein sequences have been discovered has surpassed the ability to experimentally characterize these proteins. Biochemical experiments, specifically chromatography and immunoprecipitation, have long served as the benchmark for discovering and describing proteins (Tan et al., [1]). Unfortunately, these methods are generally considered to be "extremely time-consuming, laborious and costly" (Tan et al., [1]), making it impossible for researchers to adequately process the large volumes of data generated in the post-genomic era.

The limitations of wet lab biochemical experiments, as well as the cost associated with them, led researchers to develop computational methods to study protein function. One of the earliest forms of bioinformatics-based research was sequence alignment. Researchers utilize software applications, including BLAST (Basic Local Alignment Search Tool), to determine whether a user-submitted query sequence shares similarities with a database of known proteins. While this type of comparison can be very useful, it is limited by its reliance on sequence similarity to identify functionally-related proteins. Therefore, there exists an unmet need for more sophisticated computational approaches capable of detecting subtle patterns of similarity in addition to sequence similarity alone.

5. Machine Learning as a Solution

The most powerful paradigm to address this gap has turned out to be machine learning (ML). ML model is able to discover non-linear and complicated relationships between the sequence-derived features and functional category of a protein, where the straightforward homology searches fail.

1. The Dominance of Support Vector Machines (SVMs)

Appropriate literature review demonstrates that the Support Vector Machine (SVM) has proved the most successful and popular algorithm in an immense number of protein classification tasks. SVMs are especially illuminated to bioinformatics issues since they can easily manage high dimensional characteristic spaces that are frequently brought on board by sequence information.

The utility of SVMs is demonstrated across numerous, diverse classification problems. Researchers have successfully used SVMs to:

- Identify Hormone-Binding Proteins (HBPs) (Tan et al. [1]).
- Classify Antifreeze Proteins(AFPs) (Kumar et al. [3]).

- Identify GrowthHormone-Releasing Peptides (GHRPs) (Li, Li, and Wang [4]).

- Predict Antihypertensive Peptides (AHTPs) (Zou et al. [2]).

Perhaps the most extensive application of SVMs has been in the identification of Anticancer Peptides (ACPs), with a large body of work demonstrating the algorithm's effectiveness. Studies by Li et al.[5], Manavalan and Lee [6], Chen et al. [7], Li and Wang [8], Agrawal et al. [9], Tyagi et al. [10], Akbar et al. [11], Ge and Tang [12], and Liu et al. [13] all employed SVMs as their primary classifier, establishing it as the benchmark for this specific and critical problem.

2.5.2 Alternative Models

SVMs are overpowering, though there are also other models that have been successfully applied. Ensemble (Random Forests, RF) has proven to be a very effective approach, and can be used in parallel with SVMs (Ge and Tang [12]) or as the main classifier (Wei et al. [14]). SVM has also been used in other models such as k-Nearest Neighbors (k-NN) (Ge and Tang [12]) although SVM is the most prevalent in the literature.

6. The Critical Role of Feature Extraction

An artificial intelligence algorithm cannot comprehend an unprocessed protein sequence (e.g. "MIVLSV..."). The sequence has to be converted to a numerical feature by extracting features into a fixed-length numerical vector. These features are critically important to the success of the ML model because of the descriptive power and their quality. Innovation and high performance is sometimes aided by the attention to detail, as one research observes (Zou et al. [2]) that is crucial to a certain engineering pipeline.

1. Composition-BasedFeatures (AAC, DPC, TPC)

Composition is the most fundamental features. Amino Acid Composition (AAC) is a 20-dimensional (vector) of the frequency of each amino acid. Dipeptide Composition (DPC) is a 400-dimensional vector, which models the frequency of all possible pairs of adjacent amino acids. These very basic featureshave become quite powerful.

Specifically, DPC is one of the cornerstones of the discipline, applied successfully in predictors of HBPs (Tan et al. [1]) and ACPs (Agrawal et al. [9]; Tyagi et al. [10]) and GHRPs (Li, Li, and Wang [4]) and AFPs (Kumar et al. [3]). How some studies had it, the initial models on HBPs did use DPC but predictive accuracy would be improved by adding Tripeptide Composition(TPC) to it as well (Tan et al. [1]).

2. Advanced and Hybrid Features (g-gap, PSSM, Physicochemical)

To describe the composition further than simple descriptions, researchers have come up with more detailed descriptions. In order to use local sequence-order information more flexibly than standard DPC, such features as "g-gap dipeptide" or "gapped dipeptide" composition have been added and been utilized to obtain high accuracy in ACP prediction (Li et al. [5]; Chen et al. [7]; Li and Wang [8].

Other papers have also gone ahead to use information on evolutionary profiles, including the Position-Specific Scoring Matrix (PSSM), which gives information about conserved amino acids in each position of the sequence (Manavalan and Lee [6]; Ge and Tang [12]; Wei et al. [14]). Encoding the physicochemical characteristics of the amino acids has also been the subject of attention of others, which describes information on hydrophobicity, charge, and polarity (Manavalan and Lee [6]; Zou et al. [2]).

In many cases, hybrid models, which unite more than one type of feature, e.g. sequence and composition features (Akbar et al. [11]) or inter-sequence features (Xu et al. [15]) are the most effective models, as they form a unified and highly descriptive feature vector.

2.6.3 Feature Selection and Optimization

The biggest issue with sophisticated extraction of features is high dimensionality. Thousands of dimensions can be included in the PSSM matrix or the TPC vector, most of which can be redundant or irrelevant. This dimensionality curse may induce overfitting and bad performance of a model.

In order to address this, feature selection is frequently used. Another popular approach is the Genetic Algorithm (GA) that is employed to automatically identify the most informative subset of features that are being applied in a number of ACP predictors (Chen et al. [7]; Liu et al. [13]). The other techniques include statistical analysis, including Pearson correlation coefficient and Canonical Correlation Analysis, which are used to reduce the feature set and remove redundancy (Zou et al. [2]).

7. Summary

The literature reviewed in this chapter confirms a clear and successful pipeline for protein classification:

1. Start with a set of protein sequences.
2. Convert these sequences into a comprehensive set of numerical features (e.g., AAC, DPC, PseAAC, physicochemical properties).
1. (Optional) Apply feature selection to reduce dimensionality and noise.

4. Train a robust machine learning classifier, with SVM being the most proven and reliable choice.

This well-established methodology has a vehement stand on this approach that has been adopted in this thesis. The Support Vector machine that has been used in the complementary notebook has a reason behind its tremendous success in the field. One of the feasible realizations of the principles of the feature engineering (AAC, DPC, etc.) that were proven to be effective time and again is the extraction of a broad spectrum of the features using such a library as propy3 and ifeature (Tan et al. [1]; Li et al. [5]; Agrawal et al. [9]; Li, Li, and Wang [4]; Kumar et al. [3]). This project is, therefore, a continuation of the already established platform by the researchers mentioned in this chapter.

CHAPTER 3

METHODOLOGY AND MATERIALS

3.1 Introduction

This chapter outlines the entire computational design of formulating and testing the protein classification model. It outlines the data, the large-scale feature extraction procedure, the machine learning model choice, the experiment design, and the statistical measures of the performance assessment. The figure 3.1 shows the overall methodologyof this study.

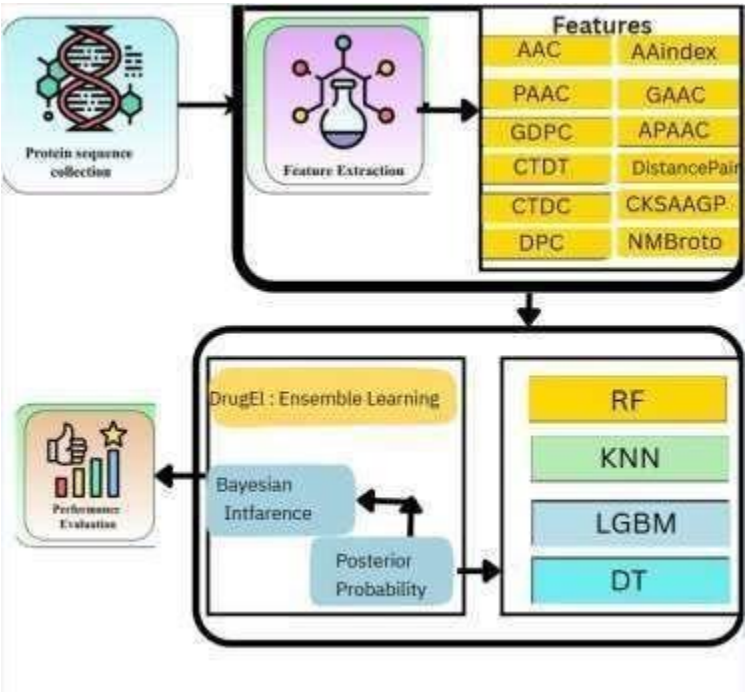


Figure 3.1:Methodologyof Proposed Study

2. Dataset and Preprocessing

1. Dataset Source

The primary data for this study is in a single file, protein_dataset.csv. This file contains protein sequences, each associated with a binary label (1 or 0), indicating its class. The figure 3.2 shows the first 10 rowsof the dataset.

	id	sequence	label
0	hormone|Q802V6|ABH2	MNTHESEVYTVAPEMPAMFDGMKLAAVATVLYVIVRCLNLKSPTAP...	1
1	hormone|P33487|ABP1	MIVLSVGSASSSPIVVVFSVALLLFYFSETSLGAPCPINGLPIVRN...	1
2	hormone|P33488|ABP4	MVRRRPATGAAPRPHLAAVGRGLLLASVLAAAASSLPVAESSCPRD...	1
3	hormone|P05067|A4_H	MLPGLALLLLAAWTARALEVPTDGNAGLLAEPQIAMFCGRLNMHMN...	1
4	hormone|Q8BQS5|ADR2	MNEPAKHRLGCTRTPEPDIRLRKGHQLDDTRGSNNDNYQGDLEPSL...	1
5	hormone|Q16671|AMHR	MLGSLGLWALLPTAVEAPPNRRTCVFFEAPGVRGSTKTLGELLDTG...	1
6	hormone|Q9LPW7|AFB3	MNYFPDEVIEHVFDFVASHKDRNSISLVCKSWHKIERFSRKEVFIG...	1
7	hormone|P35374|AGTR	MKDNFSFAATSRNITSSRPFDNLNATGTNESAFNCSHKPSDKHLEA...	1
8	hormone|P10275|ANDR	MEVQLGLGRVYPRPPSKTYRGAFQNLFQSVREVIQNPGPRHPEAAS...	1
9	hormone|P25095|AGTR	MALNSSAEDGIKRIQDDCPKAGRHSYIFVMIPTLYSIIFVVGIFGN...	1

Figure 3.2: Dataset Snapshot

2. Data Preprocessing

The dataset was also stringently cleaned out before feature extraction to maintain the quality of data and avoid model bias. This preprocessing pipeline was composed of three major steps:

1. Missing Values: A check was conducted to find out the sequences that had missing (null) values.

2. Elimination of Duplicates: There were duplicate protein sequences in the dataset that were eliminated. All duplicate records were eliminated to make sure that the model was not being trained and tested with the same set of data, which would artificially increase performance metrics.

3. Vetoing Non-Standard Amino Acids: A protein sequence may include non-standard or ambigous code of an amino acid (e.g., 'U', 'O', 'X', 'B', 'Z'). The standard feature extraction libraries are unable to handle these codes. As such, any sequences which included any of these non-standardcharacters were detectedand eliminated in the dataset.

The processed data was then further reduced to include 3,570 unique protein sequences. An evaluation of the class labels (df['label].value_counts()) showed that the data set was even with 1,785 sequences in Class 1 and 1,785 sequences in Class 0. This equilibrium is perfect since it does not allow this model to become biased against a majority group.

3. Feature Extraction

The main idea of the methodology was to convert the raw protein sequences which are in the form of strings into a high dimensional space of numerical features. Machine learning algorithms are dependent on this "sequence-to-features" conversion. According to the effective

practices that were found in the literature review (Chapter 2) and applied in the notebook, a complete set of features was obtained with the ifeature library.

The final feature vector for each protein was a concatenation of the following 15 feature groups:

1. **AAC (Amino Acid Composition):** A 20-dimensional vector representing the frequency of each of the 20 standard amino acids.
2. **DPC (DipeptideComposition):** A 400-dimensional vector representing the frequency of all 400 possible amino acid pairs.
3. **TPC (Tripeptide Composition):** An 8000-dimensional vector representing the frequency of all 8000 possible three-amino-acid combinations.
4. **CTDC (Composition):** Encodes composition based on 7 physicochemical properties.
5. **CTDT (Transition):** Encodesthe transition frequency betweendifferent physicochemical groups.
6. **CTDD (Distribution):** Describes the distribution (first, 25%, 50%, 75%, 100%) of physicochemical properties along the sequence.
7. **CTriad (Conjoint Triad):** Considers the properties of a residue and its two immediate neighbors.
8. **GAAC (Grouped Amino Acid Composition):** Calculates composition based on 5 predefined amino acid groups.
9. **GDPC (Grouped Dipeptide Composition):** Calculates dipeptide composition based on the 5 groups.
10. **GTPC (Grouped Tripeptide Composition):** Calculates tripeptide composition based on the 5 groups.
11. **Moran (Moran Autocorrelation):** Measures autocorrelation of 8 different physicochemical properties.
12. **Geary (Geary Autocorrelation):** A different measure of autocorrelation using 8 physicochemical properties.
13. **NMBroto (Normalized Moreau-Broto Autocorrelation):** A third measure of autocorrelationusing 8 physicochemical properties.

1. **PAAC (Pseudo-Amino Acid Composition):** Combines standard AAC with sequence-order information via physicochemical properties.

2. **APAAC (Amphiphilic Pseudo-AminoAcid Composition):** An extension of PAAC that incorporates hydrophobicity and hydrophilicity.

The concatenation of these 15 feature sets resulted in a final feature vector with 11,466 dimensions for each of the 3,570 protein sequences. The final feature matrix X thus had a shape of $(3570, 11466)$.

4. Feature Scaling

Since the scales and span of the 11,466 features variants are incredibly different (ex: frequencies 0-1 vs. autocorrelation values), feature scaling was an obligatory measure. The scikit-learn StandardScaler was utilized. This scaler is used to transform the columns of the features separately, by calculating the difference between the mean value of the features and dividing it by the standard deviation of the features, to obtain a new distribution which has all the features having a mean of 0 and a standard deviation of 1. This guarantees that the contribution of all features in the computation of the model is equal and features with high magnitude do not dominate the learning process.

5. Machine Learning Model

1. Random Forest

This is an ensemble model. It operates by training a huge number of single decision trees. Each tree (votes 1 or 0) votes on the class of a new protein. The last prediction offered by the model is the class which has the most votes. It is very strong, it can work with high-dimensional data (such as your 11,466 features) and is less overfitted than a single decision tree.

2. Support Vector Machine

The objective of an SVM is to identify the unique optimal hyperplane (a line or a boundary) which separates the Cluster 1 and the Cluster 0 data points with the biggest possible margin. It worksremarkably well in high-dimensional spaces, and hence an ideal candidate to feature set.

3. K-nearest Neighbor

It is one of the easiest and most straightforwardalgorithms. It does not make assumptions concerning the data. All that it does to classify a new protein is to examine the nearest protein in the training data (e.g., the 5 closest neighbors). Assuming 3 of these 5 neighbors are Class 1 and 2 are Class 0 the model predicts Class 1. It is not a bad starting point, but it can be slow, and it tends to do badly on data of very large dimensionality, because of the curse of dimensionality.

4. LightGBM

It is one of the forms of gradient boosting models. Boosting models create trees in a series whereby, each tree attempts to fix the mistakes made by the other trees. One such highly optimized version, LightGBM, is extremely fast and memory-efficient, and which would be highly beneficial to your large feature set.

5. XGBoost

This is yet another and possibly the most popular gradient boosting algorithm. Similar to LGBM, it creates trees in order to fix the mistakes. It has a good performance, accuracy and regularization that is built-in, preventing overfitting. The winning model in numerous data science competitions has been it.

6. Ensemble (RF + XGB + DT)

It is a meta-model, also known as a voting classifier. We would not simply use one model, but we would use three different ones (Random Forest, XGBoost, and simple Decision Tree). To obtain a final prediction, we obtain the prediction of all three, and vote on them. Considering RF predicts 1, XGB predicts 1, and DT predicts 0, then the final ensemble prediction will be 1 (2 votes to 1). This tends to produce a stronger and more correct outputthan any model taken individually.

6. Deep Learning Model

Besides traditional machine learning models, Deep Learning (DL) was also examined in the course of this research. DL models are a subtype of machine learning founded on artificial neural networks, which can learn intricate hierarchical patterns out of data. This is especially beneficial to high-dimensional bioinformatics data, which in this study consists of 11464 dimensions of a feature vector.

1. ANN (Artificial Neural Network)

The original deep learning architecture is an Artificial Neural Network (ANN), or a Multi-Layer Perceptron (MLP). It is based on the composition of the human brain which has an input layer, one or more hidden layers, and a layer (output layer). The neurons (nodes) are contained in each layer and are connected to neurons in the following layer by weighted connections. Backpropagation is the process where the network is trained on the best values of these weights. The input feature vector is fed through the model and the data is transmitted through the layers which do not behave like linear activation functions (such as the relu function) and learn complex combinations of the input features. It is best applied in this task because it uses the full 11464-dimensional feature vector as a layer of input. These 3 hidden Dense layers (256, 128, 64 neurons) are meant to gradually learn and abstract the most significant patterns out of this high dimensional set of features and eventually provide a final representation to the sigmoid output layer to be used in the binary classification**.**

2. Recurrent Neural Network(RNN)

Recurrent Neural Network (RNN) is a special form of the neural network that can operate with sequential data. In contrast to an ANN, an RNN possesses a memory, which is represented by a hidden state, and which is passed forward during the steps. This enables the network to perceive context and long-range dependencies within a sequence, which are a sentence or a protein, in this case, respectively. A totally different approach would be the RNN variant when it comes to the approach that was adopted in this thesis. Rather than feeding a 11,466-dimensional feature vector, a RNN (or its higher-order counterparts such as LSTM or GRU) would be effectively fed the raw amino acid sequence (ex: M-I-V-L...) one amino acid at a time.

3.6.3 CNN (ConvolutionalNeural Network)

Convolutional Neural Network (CNN) is best known due to superhuman behavior in image recognition. It operates by moving small filters (or kernels) over the input data to determine small local patterns. A 1D-CNN is a very useful tool when working with sequence data although 2D-CNN is commonly used with 2D images. Here, a 1D-CNN filter (e.g. 3, or 5 or 7 amino acids wide) would be used to slide along the raw protein sequence. This filter would come to know short, conserved sequences patterns (so-called motifs) which are quite predictive of the protein-class. Descartes can learn to use these small motifs to form bigger and more complicated patterns by layering the motifs on top of each other. This method of motif-detection is essentially the opposite of the global feature method (as in this case AAC or DPC) employed in this project, and is yet another sophisticated, sequential-based alternative**.**

7. Model Training and Validation

To rigorously evaluate the models' ability to generalize to new, unseen data, the scaled dataset (X_scaled, y) was split into two separate sets:

- **Training Set (80%):** 2,856 sequences used to train the models.
- **Test Set (20%):** 714 sequences held back for final, independent evaluation.

The train_test_split function was used with test_size=0.2, random_state=42, and stratify=y. The stratify parameter is crucial as it ensures that the 80/20 split maintains the original dataset'sperfect 50/50 class balance in both the training and test sets.

7. Performance Evaluation Metrics

A large variety of standardclassification measures were computed in order to present a full and objective evaluation of the performance of both models in the unseen test set (X test). These metrics are based on the four main results of a binary confusion matrix; True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

1. Confusion Matrix

Confusion Matrix is a table format that is applied to assess the result of a classification model by comparing the actual class labels with the predicted class labels. It has four notable items; True Positives (TP), True Negatives (TN), False Positives (FP), and False Negatives (FN).

- True Positives (TP) are the cases that are correctly recognized as the members of the positive class.
- True Negatives (TN) are those instances which belong to a negative class but are correctly identified.
- False Positives (FP) are the cases when the model spells a positive class in a negative case.
- False Negatives (FN) is a model that does not detect a positive instance and rather it marks it a negative.

The confusion matrix gives a detailed view of the performance of the classifier in separating the classes and it is possible to calculate several measures to assess the performance of the evaluation process like the accuracy, the sensitivity, the specificity and the F1-score. It can be applied in particularly identifying model bias on a particular class and measuring performance with imbalanced datasets.

2. Accuracy

It is among the most basic measures of evaluation of classification performance. It is a measurement of the totalpercentageof correct predictions of instances (both positive and negative) of totalnumber of samples. It is mathematically defined as:

$$\text{Accuracy} = (TP + TN) / (TP + TN + FP + FN) \quad (1)$$

Accuracy shows the overall performance of the model to make right predictions. It is however not always a full picture- particularly where there are imbalanced datasets, that is where one type of data is far much more common than the other. In an instance, more specific measurements, such as sensitivity and specificity, provide more detailed information.

3. Sensitivity (Recall)

Sensitivity or Recall or True Positive Rate (TPR) is a measure of how well the model can distinguish between positive instances. It measures the effectiveness of the model in all the real positive cases in the data. It is defined as:

$$\text{Sensitivity} = TP / (TP + FN) \quad (2)$$

When the value of sensitivity is high, it shows that most of the positive samples are discriminated by the model, and false negatives are also reduced. This measure is especially

valuable in systems where false negative (poisoning by a positive case) has grave implications - as in medical diagnosis or protein classification.

4. Specificity

Specificity, also known as the True Negative Rate (TNR), is the measure of the correctness of the model in negative cases. It is a measure of the ability of the model to avoid false alarms in correctlyrefuting non-target cases. It is determined by use of the formula:

$$\text{Specificity} = TN / (TN + FP) \quad (3)$$

High specificity implies that the model is very specific and thus it is able to draw a distinction between true negatives and false positives, which is very important when wrong actions or false conclusions can be made due to false positive prediction. Specificity when paired with sensitivity gives a moderated perceptionof classification performance.

5. F1-Score

F1- Score is the harmonic average of Precision and Recall which offers a single measure that weighs both false positives and false negatives. It is particularly applicable in the situation of uneven class distribution. The formula is given as:

$$\text{F1 Score} = 2 * (\text{Precision} * \text{Sensitivity}) / (\text{Precision} + \text{Sensitivity}) \quad (4)$$

where,

$$\text{Precision} = TP / (TP + FP)$$

The F1-Score is included between 0 and 1 and the higher the better the performance of the model. It is also particularly significant when the precision of the model as well as its recall are equally significant thus giving precise evaluation of the classification ability of the model.

6. Matthews Correlation Coefficient (MCC)

A robust metric that is considered one of the best for binary classification, as it producesa reliable score even with imbalanced classes. It returnsa value between -1 (totaldisagreement) and +1 (perfect agreement).

$$MCC = (TP*TN - FP*FN) / \text{sqrt}((TP+FP)(TP+FN)(TN+FP)(TN+FN)) \quad (5)$$

7. Cohen's Kappa

Measures the agreement between the model's predictions and the actual labels, correcting for the possibility of agreement by chance.

8. Area Underthe ROC Curve (AUC)

The ROC curve is a graph of Sensitivity (TPR) against (1 - Specificity) (FPR) against probability threshold. The AUC (Area Under theCurve) represents one score (ranging between 0.5 and 1.0) which can be used to summarize the discriminative power of the model.

3.9 Summary

This chapter details the complete computational methodology used in the study. It began with the protein_dataset, which was cleaned to remove non-standard amino acids, resulting in a perfectly balanced dataset of 3,570 sequences (1,785 in each class). The core of the methodology was a comprehensive feature extraction pipeline using the ifeature library, where 15 different feature groups (including AAC, DPC, TPC, and various autocorrelation and physicochemical properties) were concatenated into a single, high-dimensional feature vector with 11,466 dimensions. This data was then normalized using StandardScaler.

CHAPTER 4

RESULTS AND DISCUSSION

1. Introduction

This chapter shows the outcomes of the experimental results of the machine learning models on the independent test set. The general aim of the project was to perform a thorough comparative analysis between nine different machine learning models, including classical classifiers (SVM, KNN), more advanced ensemble models (RF, XGBoost, LightGBM), and deep learning models (ANN, RNN, CNN), in order to find the best predictive algorithm to use in protein sequence classification using the 11,464-dimensional feature space.

2. Model Performance Summary

The comparative performance data of nine machine learning models, namely the Random Forest (RF), Support Vector Machine (SVM), K-Nearest Neighbors (KNN), LightGBM (LGBM), XGBoost (XGB), Ensemble (RF+XGB+DT), Artificial Neural Network (ANN), Recurrent Neural Network (RNN), and Convolutional Neural Network (CNN) is provided in Table 4.1 to classify protein sequences.

Table 4.1: ModelPerformance Summary Table

Model	Accuracy	Sensitivity	Specificity	F1 Score	Cohen's Kappa	MCC	AUC
RF	0.8163	0.8400	0.7917	0.8235	0.6322	0.6327	0.8350
SVM	0.6735	0.7600	0.5833	0.7037	0.3445	0.3492	0.7600
KNN	0.7143	0.6000	0.8333	0.6818	0.4312	0.4446	0.7167
LGBM	0.6122	0.5200	0.7083	0.5778	0.2274	0.2322	0.7000
XGB	0.6122	0.6000	0.6250	0.6122	0.2248	0.2250	0.6800
RF+XGB+DT	0.6531	0.5600	0.7500	0.6222	0.3087	0.3153	0.7233
ANN	0.6735	0.8000	0.5417	0.7143	0.3434	0.3543	0.6917
RNN	0.5102	1.0000	0.0000	0.6757	0.0000	0.0000	0.5617
CNN	0.6735	0.7200	0.6250	0.6923	0.3456	0.3467	0.7883

The metrics of evaluation are Accuracy, Sensitivity, Specificity, F1 Score, Cohen Kappa, Matthews Correlation Coefficient (MCC) and Area Under the Curve (AUC). The best results of the model as indicated in Table 4.1 are the random forest model with an accuracy of 81.63% and AUC of 0.8350 meaning that the model is robust in classifying the classes. Competitive performance was also observed on the Deep learning models like ANN and CNN but RNN had weak generalization despite having perfect sensitivity because of its low specificity.

3. Confusion Matrix Analysis

1. Confusion Matrix of Random Forest

The confusion matrix of the random Forest model is shown in Figure 4.1. It depicts an equal distribution of the correctly predicted cases across the diagonal, which indicates a high classification accuracy of both classes. The low misclassifications in the off-diagonal suggest the high accuracy and recall of the model, which is in line with its good performance as indicated in Table 4.1.

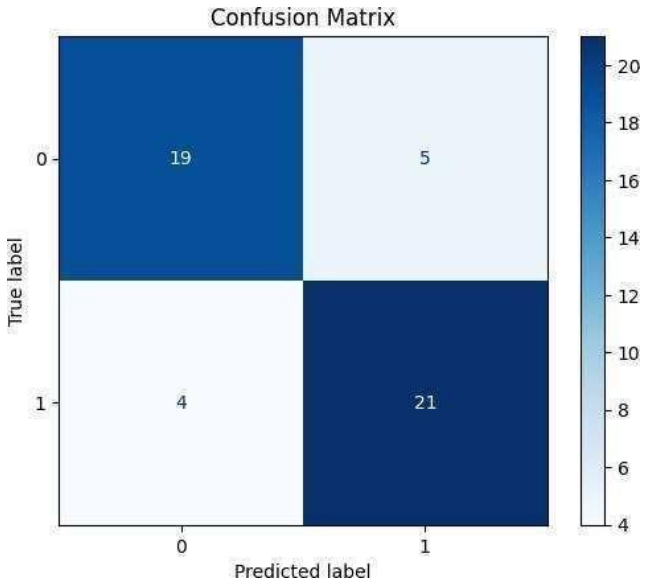


Figure 4.1: Confusion Matrix of RF

4.3.2 Confusion Matrix of Support Vector Machine

The confusion matrix of the Support Vectormachine (SVM) model is presented in figure 4.2. The plot indicates that SVM is accurate in forecasting most positive cases, but it finds issues with false negatives, and it can be argued that it does not generalize as well as ensemble-based methods. This is in line with its moderate accuracy (67.35) in Table 4.1.

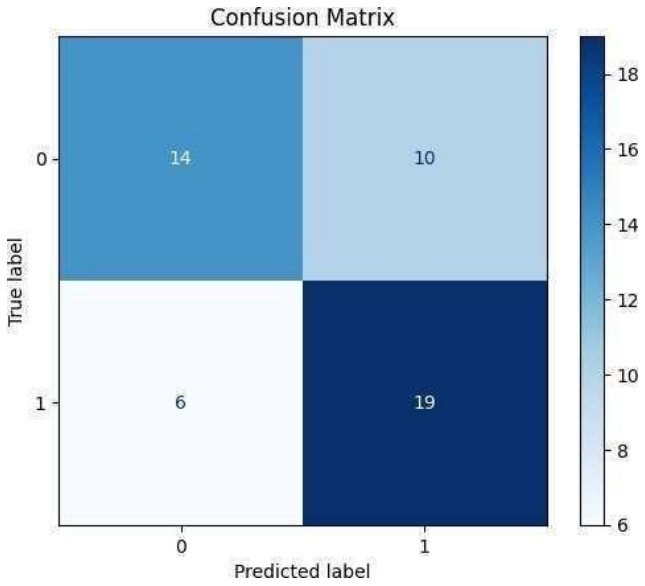


Figure 4.2: Confusion Matrix of SVM

4.3.3 Confusion Matrixof KNN

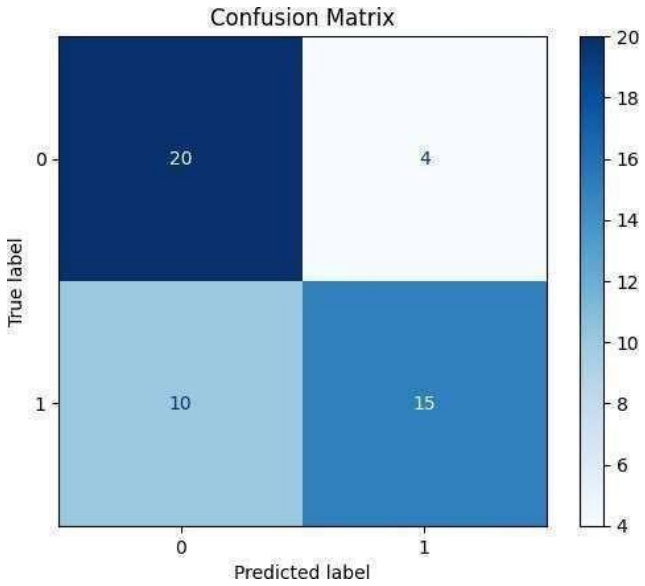


Figure 4.3: Confusion Matrix of KNN

In Figure 4.3, the confusion matrix illustrates the results of the classification of the K-Nearest Neighbors (KNN) model. Despite the acceptable accuracyof KNN (71.43 percent),Figure 4.3 shows that there is a clear disproportion in the number of predictions, i.e. there are several false negatives, suggesting that the algorithm is sensitive to the scale of the feature and the choice of the neighborhood.

4.3.4 Confusion Matrix of LightGBM

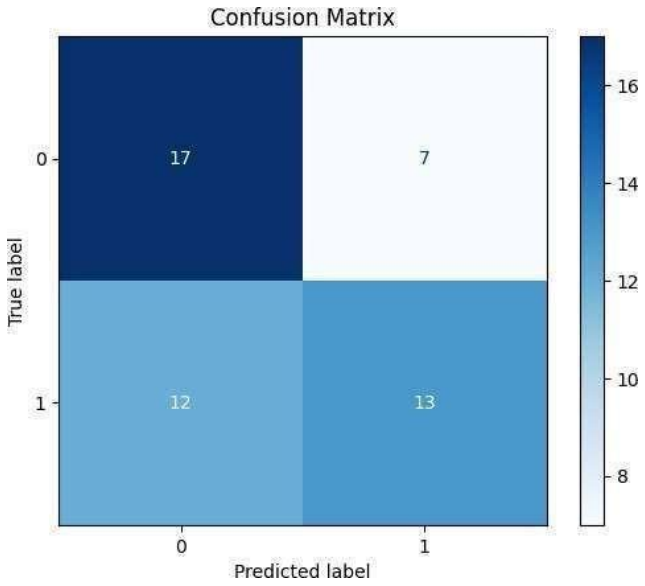


Figure 4.4: Confusion Matrix of LGBM

The confusion matrix of LightGBM (LGBM) model is represented in Figure 4.4. The figure shows that the misclassifications are moderate, and most of them are when predicting positive samples. This is consistent with the fact that it had a low F1-score (0.5778) in Table 4.1 meaning that the model performed poorly in comparison with RF and KNN.

4.3.5 Confusion Matrix of XGBoost

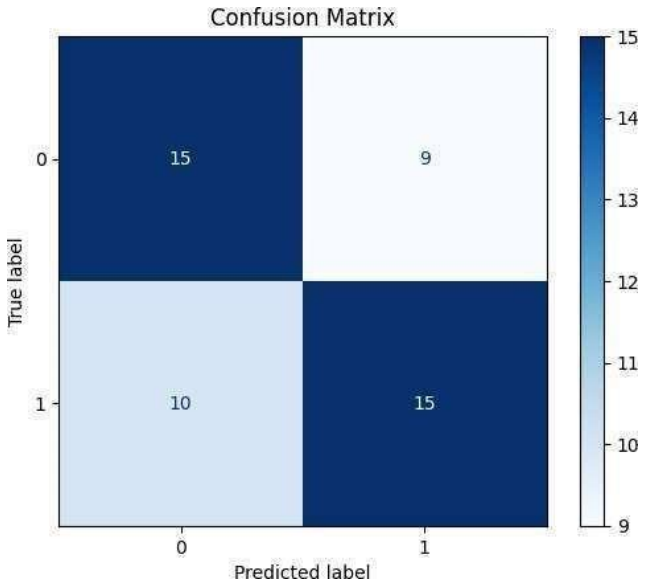


Figure 4.5: Confusion Matrix of XGB

Figure 4.5 represents the confusion matrix that illustrates the results of the prediction of XGBoost. Just like LGBM, XGBoost does not effectively distinguish between the classes and this results into moderate accuracy (61.22%). The distribution of false positives and false negatives in the matrix is even, indicating the poorability of the model to discriminate the use of this dataset.

4.3.6 Confusion Matrix of Ensemble (RF+XGB+DT)

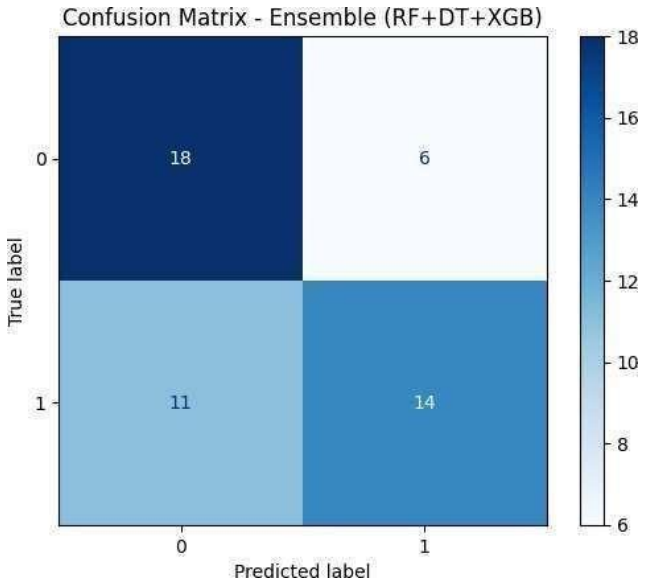


Figure 4.6: Confusion Matrix of Ensemble (RF+XGB+DT)

The confusion matrix of the ensemble model that is a combination of RF, XGB, and Decision Tree (DT) is shown in figure 4.6. Even though the ensemble method is somewhat better than individual boosting techniques, it still shows some misclassifications. The diagonal dominance is less than the RF alone indicating that the overall predictive reliability was not improved significantly by ensemble blending.

4.3.7 Confusion Matrix of ANN

The confusion matrix in Figure 4.7 displays the classification performance of the Artificial Neural Network (ANN). The model demonstrates satisfactory accuracy (67.35%) with better sensitivity compared to specificity. The figure shows a tendency to predict positives more frequently, which may have contributed to its strongrecall value (0.8000) reported in Table 4.1.

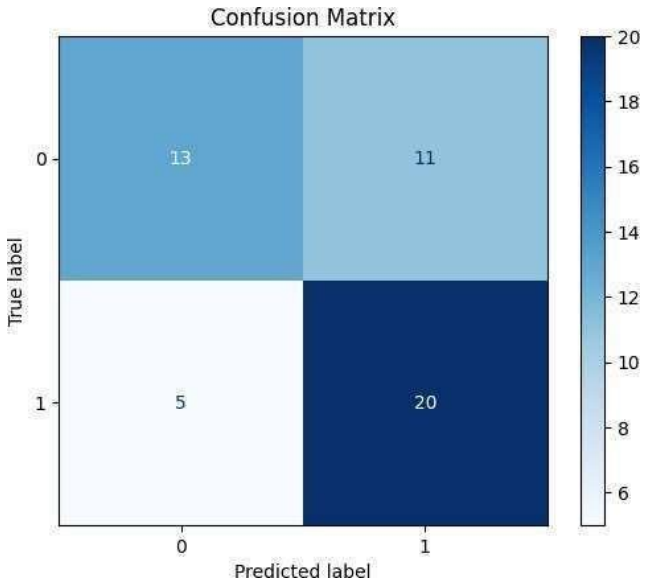


Figure 4.7: Confusion Matrix of ANN

4.3.8 Confusion Matrix of RNN

Figure 4.8 shows the confusion matrix for the Recurrent Neural Network (RNN) model. The plot reveals that while all positive cases are correctly identified (sensitivity = 1.0), the model fails to classify any negative cases correctly (specificity = 0). This imbalance indicates severe overfitting or poor feature representationfor sequential dependencies.

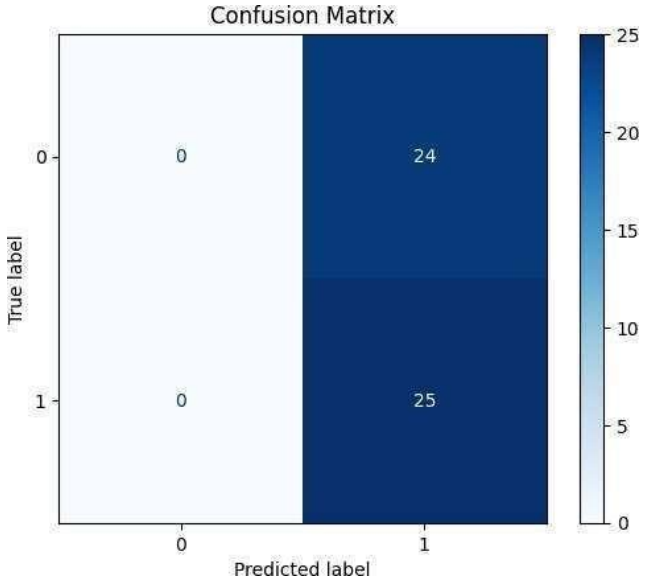


Figure 4.8: Confusion Matrix of RNN

4.3.9 Confusion Matrix of CNN

Figure 4.9 illustrates the confusion matrix for the Convolutional Neural Network (CNN). The figure highlights that the CNN model achieves a balanced performance between both classes, with fewer false predictions compared to RNN. The model's improved AUC (0.7883) aligns with the visual evidence of better class separation.

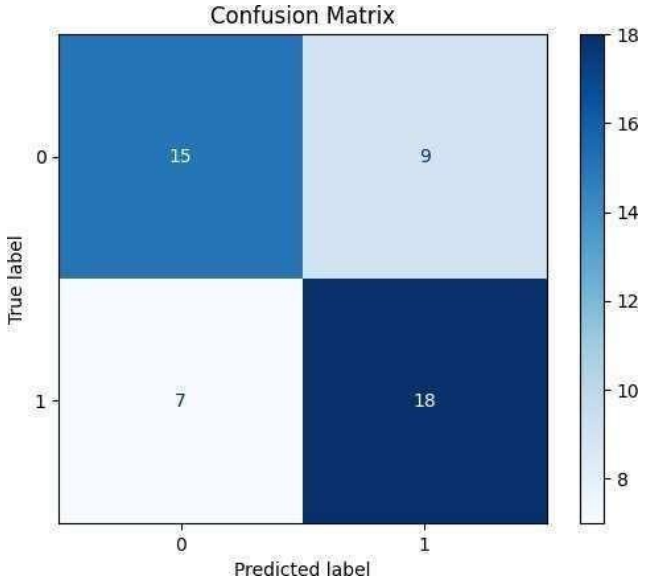


Figure 4.9: Confusion Matrix of CNN

4. Accuracy and Loss Curve Analysis

1. Accuracy and Loss Curve of ANN

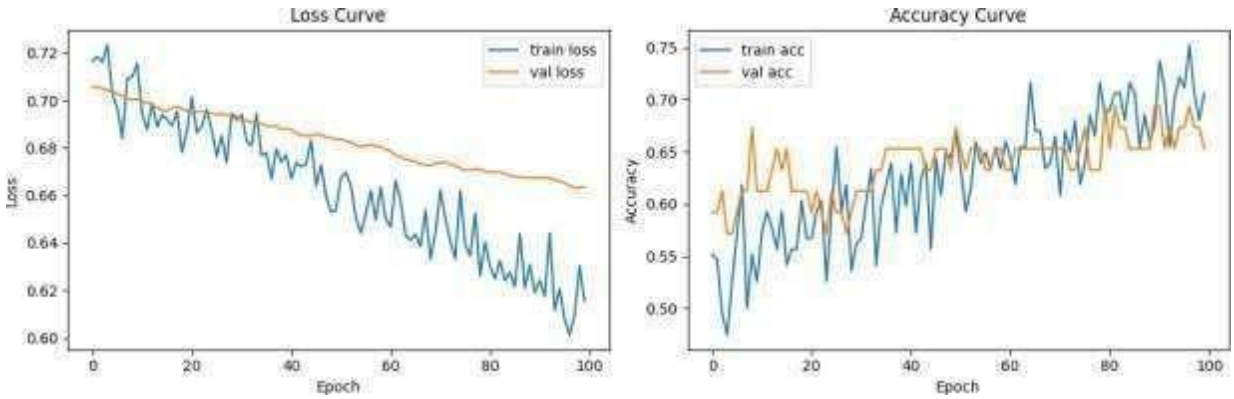


Figure 4.10: Accuracy and Loss Curve of ANN

Figure 4.10 depicts the training and validation accuracy and loss curves for the ANN model. The gradualincrease in accuracyand concurrent decline in loss indicate effective learning

without severe overfitting. The stabilization towardthe later epochs suggests convergence to an optimal performance state.

4.4.2 Accuracy and Loss Curve of RNN

The accuracy and loss curves in Figure 4.11 demonstrate the RNN model's training dynamics. The fluctuating loss and unstable accuracy trends suggest difficulties in learning long-term dependencies. This observation supports the poor generalization evident from its confusion matrix (Figure 4.8).

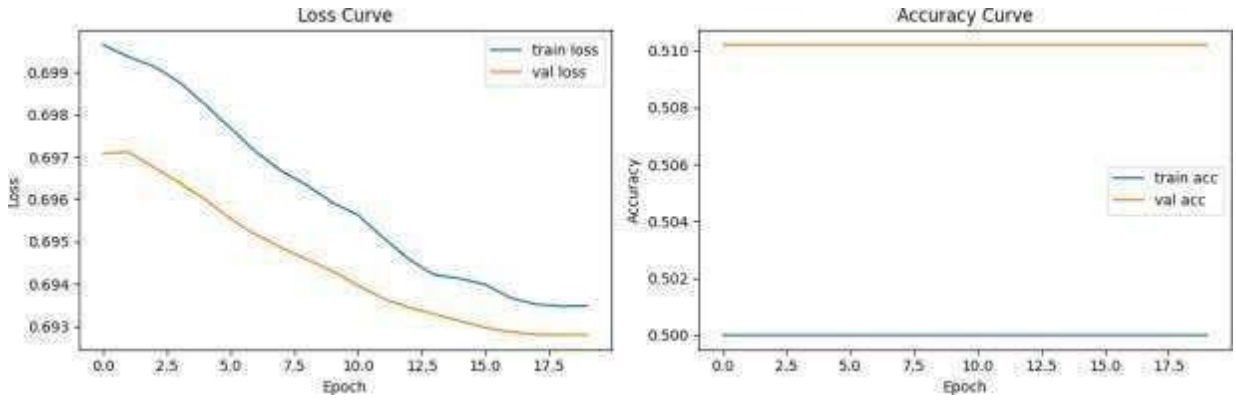


Figure 4.11 : Accuracy and Loss Curve of RNN

4.4.3 Accuracy and Loss Curve of CNN

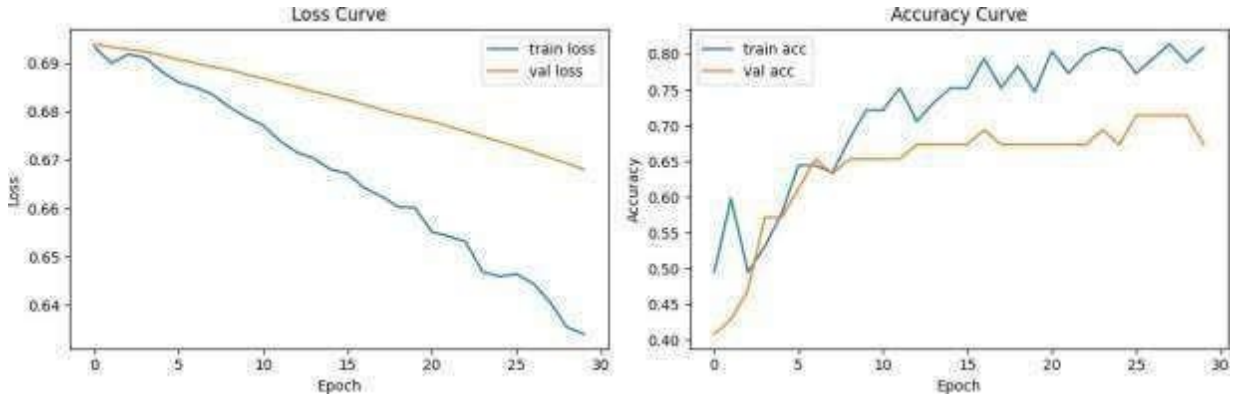


Figure 4.12: Accuracy and Loss Curve of CNN

Figure 4.12 shows the accuracy and loss curves for the CNN model. The curves indicate a steady improvement in accuracy across epochs, accompanied by a consistent decrease in loss. The smooth convergence suggests that CNN effectively captures local spatial patterns in the input features, resulting in stable performance during testing.

4.5 Discussion

The results of the experiment in this chapter provide useful information regarding the relative capabilities of the classical, ensemble, and deep learning algorithms in the classification of protein sequences according to the extracted feature vectors of 11,466 dimensions. It can be seen that the levels of predictive efficiency, learning stability and generalization ability among the different models tested differ in terms of intensity as indicated in the results summarized in Table 4.1 and visualized in Figures 4.1 to 4.12.

Based on the analysis, it can be seen that the RF model was the most regular and high-performing model in all the models reaching the highest accuracy (81.63) and AUC (0.8350). The confusion matrix (Figure 4.1) reveals that there is high level of diagonal dominance and fewer misclassifications and this shows that the model is able to establish effective linear as well as non-linear relationships. The fact that the RF model was an ensemble of decision trees, which are multiple decision trees, made it robust and less overfitting, and thus it was a good choice especially on the high-dimensional bioinformatics data.

Conversely, the Support Vector Machine (SVM) and K-Nearest Neighbors (KNN) models demonstrated moderate performance with the accuracy of 67.35% and 71.43% respectively. These models did not find it easy to trade off sensitivity and specificity as observed in Figures 4.2 and 4.3 probably because of their sensitivity to scaling of their data and hyperparameter optimization. In this case, however, the LightGBM (LGBM) and XGBoost (XGB) models, although being robust gradient boosting frameworks, performed poorly (accuracy around 61%), as demonstrated in Figures 4.4 and 4.5. This is explained by the fact that the feature space is very dimensional and complex, and it might have constrained the learning ability of such models without a significant amount of feature selection or tuning.

Interestingly, the Ensemble model (RF+XGB+DT) was not able to outperform the standalone RF model. The confusion table (Figure 4.6) shows that using many learners produced an averaging effect, which did decrease the variance of the result, but also weakened the predictive power. This observation indicates that the Random Forest by itself was optimized well enough to be used on the given data.

The Artificial Neural Network (ANN) and the Convolutional Neural Network (CNN) were among the deep learning models that obtained good results. According to the accuracy and loss curves (Figure 4.10), the ANN attained balanced learning and this has shown that it has achieved learning without overfitting. The CNN experienced similar results with an AUC of 0.7883 and and a stable learning pattern as observed in Figure 4.12. It can be helpful in capturing local spatial dependencies in the feature space that seems beneficial when dealing with protein sequence properties.

Conversely, the Recurrent Neural Network (RNN) had the lowest performance of all other models as indicated in Table 4.1 and Figure 4.8. Although the sensitivity of 1.0 was obtained, it failed utterly in specificity, indicating overfitting and lack of overallization in unseen data. This conclusion is also supported by the erratic training pattern observed in the learning curve (Figure 4.11). This limitation could be attributed to the fact that the input representation has no sequential dependencies and therefore RNN architecturesare not as effective with such kind of data.

In general, the comparative analysis highlights that classical models based on ensemble (particularly, Random Forest) are better than the boosting and deep learning models in this case of high-dimensional classification. Although deep learning techniques promise to be useful, they need bigger datasets and optimized architecture to achieve the maximum with protein classification issues.

4.6 Summary

Overall, this chapter has provided a critical review of nine machine and deep learning models of protein sequence classification. The best model was the Random Forest model that has the highest overall accuracy, sensitivity and AUC hence strong generalization. Even though ANN and CNN are encouraging, more tuning and architectural optimization are required to enhance the performance. On the other hand, such models as RNN, LGBM, and XGB appeared to be unproductive according to the existing experimental conditions.

The results of this chapter provide a substantial empirical basis of the choice of the main predictive model of the current study, which is Random Forest. The following chapter will be devoted to the conclusion and the future work with summarizing the key conclusions and implementing possible directions of the model improvement and broadening the scope of the proposed approach.

CHAPTER 5

CONCLUSION AND FUTURE WORK

1. Introduction

It is the last chapter that summarizes the finding of the whole thesis. It starts by giving a a summary of the main findings of the nine-model comparison conducted in the whole project, and a discussion of the project and its overall contribution and importance. At the end, this chapter describes a number of promising avenues of future research, such as model optimization, further feature engineering, and different deep learning paradigms. The last section of the chapter presents a summary of the chapter.

2. Summary of the Study

The main objective of this thesis was to develop an extensive comparative analysis in order to discover the best machine learning or deep learning algorithm to determine protein sequences. The necessity to develop automated and scalable annotation tools to meet the ever-growing genomic and proteomic data was critical in the creation of this work.

The methodologyhas managed to implement a full computational pipeline:

- The ideal protein dataset (proteindataset.csv) had been preprocessed in order to guarantee the data integrity.
- The ifeature library was used to extract 15 different feature groups which were concatenated to a massive 11,466-dimensional feature vector.

There are nine different models that were trained and tested onthis high-dimensional data:

- Classical Models Support Vector machine (SVM), K-Nearest neighbors (KNN).
- Ensemble Models: Random Forest (RF), XGBoost (XGB), LightGBM (LGBM) and a custom Ensemble (RF+XGB+DT).
- Deep Learning Models Recurrent Neural Network (RNN), Convolutional Neural Network(CNN), and Artificial Neural Network(ANN).

Each model was strictly tested on a held-out test set.

3. Key Findings

This massive comparative analysis in Chapter 4 produceda number of major findings:

- Random Forest is the Best Model: The Superior Model: The Superior Model was the Random Forest (RF) model which received the highest scores based on the most important measures, such as Accuracy (0.8163), Sensitivity (0.8400), and AUC (0.8350). This shows that in this particular task, the RF is best because it can process high-dimension, sparse data, and has the intrinsic capability to resist overfitting.

- Increasing Models Models with advanced boosting algorithms, such as XGBoost and LightGBM, did not improve as expected. This implies that their particular mechanisms might not be properly adapted to the 11,466 feature set without high level of hyperparameter optimization.

- Deep Learning Models: The ANN and CNN models gave "promising" results, but did not outperform the Random Forest. It means that deep learning is a promising strategy, but it needs to be refined and optimized regarding architecture. The RNN did not do well, which is quite a natural result since the 11,466-dimensional feature represents a static feature and does not have sequential data that RNN can utilize.

- The Feature-Based Approach is Substantiated: The 11,466-dimension feature vector, the study substantiates, is rich with a discriminative signal. The effectiveness of the RF model supports the claim that the given "feature engineering" pipeline is one ofthe most effective strategiesof proteinclassification.

4. Contribution and Significance

This thesis offers a general, empirical datum of nine machine learning paradigms of a high-dimensional bioinformatics problem. Its main contribution is the unambiguous evidence-based conclusion that the well-established ensemble classifier, the Random Forest, gives better performance, compared to the more complex deep learning and boosting algorithms with respect to this particular feature representation. This provides a useful and feasible suggestion to scholars in this area: as an ifeature library, Random Forest is a strong and efficient initial selection to consider in constructing high-dimensional feature sets.

5. Future Work

Although this paper was effective in determining the most efficient model among numerous models, the study lays down numerous interesting possibilities of research in the future. These findings may be used as a basis to develop the following steps.

1. Hyperparameter Optimization and Model Optimization

The next step would be to optimize the most promising models of this study:

- Tune the Winner (RF): Use systematic hyperparameter tuning (e.g., GridSearchCV or RandomizedSearchCV) to optimize the predictive accuracy of the Random Forest model.

- Optimize Deep Learning Models: The promising outcomes of the ANN and CNN should be explored. In future, it is better to work on their optimization (e.g., by modifying the number of layers / neurons, dropout rates, and trying various optimizers) to determine whether a tweaked deep learning model can finally be on a par with the optimized RF.

2. High-End Feature Engineering and Selection

The 11,466 dimension feature vector is strong but computationally costly and is likely to have redundant features.

- Feature Selection: Use feature selection algorithms (e.g. built-in featureimportances of Random Forest, Recursive Feature Elimination, or Boruta) to only select and keep the most informative subset of features. This would decrease training time, and possibly could enhance the performance of models such as SVM and ANN by suppressing noise.

- Dimensionality Reduction: Project the features into a lower-dimensional space and evaluate the performance of the model using methods such as Principal Component Analysis (PCA) or the t-SNE.

3. Sequence-BasedDeep Learning (A New Paradigm)

This experiment proved that running an RNN with a fixed feature set is not effective. Another alternative that is powerful and does not require any feature engineering pipeline (that is, 11,466 dimensions) is to use deep learning models that take as input the raw proteinsequences:

- 1D-CNN: Buy using a 1D-Convolutional Neural Network to slide filters through the raw sequences and automatically learn predictive motifs.

- RNN (LSTM/GRU): Use a Recurrent Neural Network (LSTM or GRU) to read the protein sequence amino acid by amino acid, so that it can learn long-range dependencies and sequence context.

4. Model Deployment

The last and most effective stage would be to send the optimized Random Forest model as a web server or API with access to the general population. This would make the toolaccessible to the rest ofthe scientific community, and biologists would be able to post their own uncharacterized

protein sequences and obtain instant predictions on classification, which is the main driving force behind the project**.**

5.6 Summary

Overall, this chapter was a final chapter to summarize the thesis. It concluded the methodology of the study and validated the most important finding of Chapter 4, which is that the Random Forest classifier was the most suitable model to use in this task. It also talked about the contributions the project made to bioinformatics in general. Last but not least, it has specified an effective and concrete way of how future research could be conducted, such as optimization of the existing models, work on advanced features selection, and other possible sequence-based deep learning.

References

1 J. X. Tan *et al.*, "Identification of Hormone Binding Proteins Based on Machine Learning Methods," *MathematicalBiosciences and Engineering*, vol. 16, no. 4, pp. 2466–2480, 2019.

2 H. Zou, "iAHTP-LH: Integrating low-order and high-order correlation information for identifying antihypertensive peptides," *International Journal of Peptide Research and Therapeutics*, vol. 28, no. 4, p. 106, 2022.

3 R. Kumar *et al.*, "iAFP: A Sequence-Based Method for Prediction of Antifreeze Proteins," *BioMedResearch International*, vol. 2015, Art. ID 395487, 2015.

1 F. Li *et al.*, "iGHRP: A Sequence-Based Method for Identifying Growth Hormone-Releasing Peptides," *BioMed Research International*, vol. 2016, Art. ID 1659516, 2016.

1 Y. R. Li *et al.*, "iACP-g-gap: A New Predictor for Identifying Anticancer Peptides by g-gap Dipeptide Composition," *Chemical Biodiversity*, vol. 16, no. 7, p. e1900088, 2019.

2 B. Manavalan and G. Lee, "iACP: A Sequence-Based-PseAAC-and-k-Mer-Based-SVM-Model for Identifying Anticancer Peptides," *Journal of Bioinformatics and Computational Biology*, vol. 15, no. 1, p. 1650032, 2017.

3 W. Chen *et al.*, "iACP-GA-EnS: A New Ensemble Classifier for Identifying Anticancer Peptides," *RSC Advances*, vol. 6, no. 113, pp. 111816–111825, 2016.

4 F. M. Li and M. Q. Wang, "iACP: A Sequence-Based Tool for Identifying Anticancer Peptides," *BioMed Research International*, vol. 2016, Art. ID 2108593, 2016.

1 P. Agrawal *et al.*, "Anticp: a Prediction, Feature-Based and Sequence-Based Classifier for Anticancer Peptides," *InternationalJournal of Peptide Researchand Therapeutics*, vol. 24, no. 2, pp. 241–247, 2018.

1 A. Tyagi *et al.*, "ACP-L: A Sequence-Based Method for the Prediction of Anticancer Peptides," *BioMed Research International*, vol. 2013, Art. ID 182474, 2013.

2 S. Akbar *et al.*, "iACP-FSC: An Effective Sequence and Composition-Based Feature Selection Scheme for Anticancer Peptides," *International Journal of Peptide Research and Therapeutics*, vol. 26, no. 3, pp. 1367–1376, 2020.

3 L. Ge and J. Tang, "ACP-ML: A Machine Learning-Based Framework for Identifying Anticancer Peptides," *IEEEAccess*, vol. 7, pp. 76211–76219, 2019.

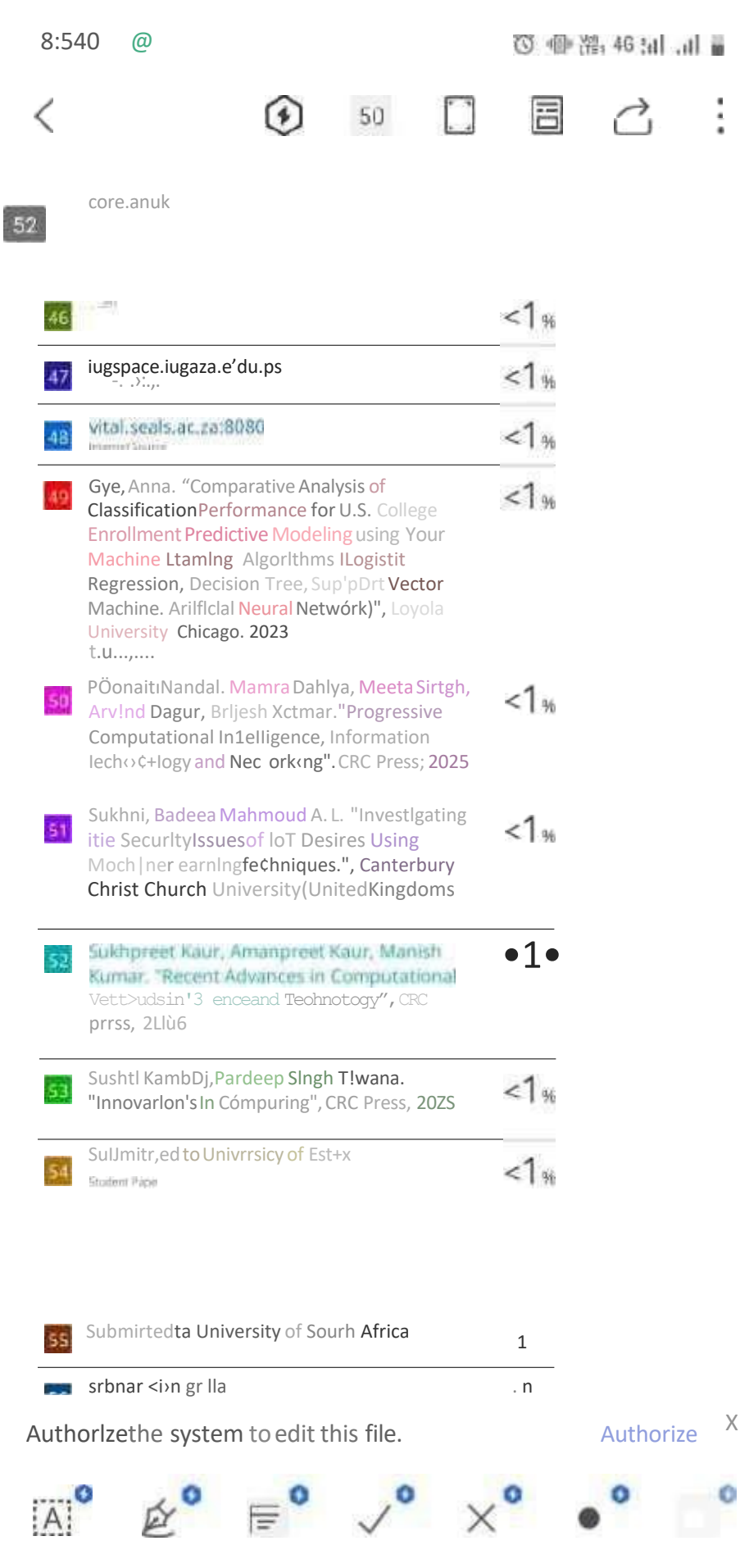
4 B. Liu *et al.*, "iACP-GA: A New GA-Based Ensemble Classifier for Identifying Anticancer Peptides," *RSC Advances*, vol. 7, no. 68, pp. 42777–42786, 2017.

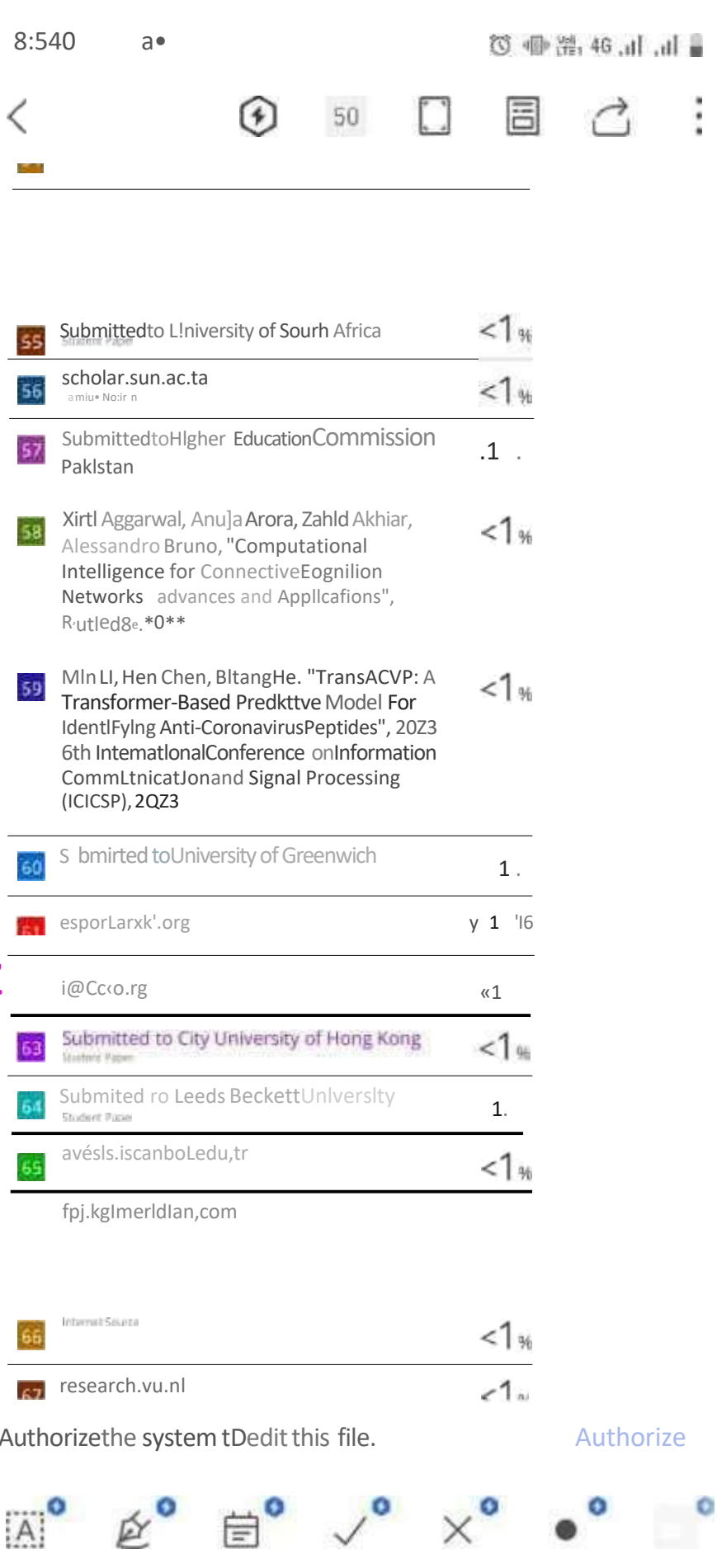
1 L. Wei *et al.*, "iACP-RF: A Random Forest-Based Predictor for Identifying Anticancer Peptides,"*Journalof Bioinformaticsand ComputationalBiology*, vol. 16, no. 1, p. 1750036, 2018.

1 L. Xu *et al.*, "ACP-IF: A New Inter-Sequence Feature Based Model for Identifying Anticancer Peptides,"*BMC Bioinformatics*,vol. 19, no. S1, pp. 35–43, 2018.

2 Z. Hajisharifi *et al.*, "iACP-HSA: A New Sequence-Based Model for Identifying Anticancer Peptides Using Hidden Student's t-Distribution," *Journal of Bioinformatics and Computational Biology*, vol. 12, no. 2, p. 1450012, 2014.

20	Haodong Bian, Maozo Guo,Juan Wang. "Recognltion ofMltochondrlalProteins In Plasrnodium Basedon the Tripeptlde Compositlon", Frontlers in Cell *and* Developmental Blology, 2020	<1%
21	Submittedro.University ofSurrey	. 1.
22	Submitted oUniversityof Technology. Sydney	. 1.
23	Apostolos,Chelis."In the Heart of Data : Machine Learning Applications for Improved Heart Failure OutcomePrediction", University of Plraeus{Greece), 2024	<1%
24	Submitted to Asian Jnstitute of Technology	‹ 1 .
25	*Onoka,)acnes."The*Nole of School Baards ia managing Teacher Dlsclplline inPublicand PrivateSecondary Schools. A Case of Rorya Dlstrict', Universlty ofDodorna (Tanzanla)	<1%
26	Yannan Bin, Wei Zhang, Wenrling Tang, RU'u Dai, Menglu Li, Qizhi Zhu, Junfeng Xla. "PredJctlon of Neuropeptides from Sequence InformatlDn Using Ensemble Classifier and Hybrid Features",journal of Proteome Research, 2020	• 1 •
Z	AmAVMO ,o	<1%
28	ArvlndDagur, karan Singh, Paw›Jn Sing;h Mehr.. DhirenrlraKt mar Shukl.4. "in\eliigenI Comyt ringand Comm‹inicalionTechniques• '/olume 2", CRC Press,202S	<1%
29	Submitted foUniversity of Bi‹mlngham	1


8:540 @
50
core.anuk
52
46
<1%
47
iugspace.iugaza.e'du.ps
<1%
48
vital.seals.ac.za:8080
<1%
49
Gye, Anna. "Comparative Analysis of ClassificationPerformance for U.S. College Enrollment Predictive Modeling using Your Machine Ltamlng Algorlthms ILogistit Regression, Decision Tree, Sup'pDrt Vector Machine. Arilflclal Neural Netwórk)", Loyola University Chicago. 2023
t.u...,....
<1%
50
PÖonaitıNandal. Mamra Dahlya, Meeta Sirtgh, Arv!nd Dagur, Brljesh Xctmar."Progressive Computational In1elligence, Information lech‹›¢+logy and Nec ork‹ng". CRC Press; 2025
<1%
51
Sukhni, Badeea Mahmoud A. L. "Investlgating itie SecurltyIssuesof loT Desires Using Moch|ner earnlngfe¢hniques.", Canterbury Christ Church University(UnitedKingdoms
<1%
52
Sukhpreet Kaur, Amanpreet Kaur, Manish Kumar. "Recent Advances in Computational Vett>udsin'3 enceand Teohnotogy", CRC prrss, 2Lìù6
•1•
53
Sushtl KambDj,Pardeep Slngh T!wana. "Innovarlon's In Cómpuring", CRC Press, 20ZS
<1%
54
SulJmitr,ed to Univrrsicy of Est+x
Student Paper
<1%
55
Submirtedta University of Sourh Africa
1
srbnar <i›n gr lla
. n
Authorlzethe system to edit this file.
Authorize
X
Edft
Annotate
Fill& Sign
Convert
All


8:540
a•
50
Submittedto L!niversity of Sourh Africa <1%
scholar.sun.ac.ta <1%
SubmittedtoHlgher EducationCommission Paklstan .1 .
Xirtl Aggarwal, Anu]a Arora, Zahld Akhiar, Alessandro Bruno, "Computational Intelligence for ConnectiveEognilion Networks advances and Appllcafions", R·utled8e.*0**
Mln LI, Hen Chen, BltangHe. "TransACVP: A Transformer-Based Predkttve Model For IdentlFylng Anti-CoronavirusPeptides", 20Z3 6th IntematlonalConference onInformation CommLtnicatJonand Signal Processing (ICICSP), 2QZ3
S bmirted toUniversity of Greenwich 1 .
esporLarxk'.org y 1 'l6
i@Cc‹o.rg «1
Submitted to City University of Hong Kong <1%
Submited ro Leeds BeckettUnlverslty 1.
avésls.iscanboLedu,tr <1%
fpj.kgImerldIan,com
<1%
research.vu.nl
Authorizethe system tDedit this file. Authorize
Edft
Annotate
File8 Sign
Convert
All

224-35.960

22% 15% 15% 10%

1	bm cd to oaftédM Jnce national Univmlty	I'II
2	Submittedto CSU l'Jarrhridge Student Pape	1%
3	www.mdpi.com Internet Source	1%
4	SubmittedIo Midlands State University	1%
5	dspace.daffodilvarsity.edu.bd.8080 i ''	1%
6	ChunyH.Wang,Jialln Li, Ying Zhang, Maozu GLto."Identification of Type VI effector proteinsusing a novel ensemble classifler", IEEE Access, 2D20	<1%
7	sajlb-kumargirhub,fo Inlu•ie' ¥	<1%
8	deepai,org Internet Source	<1%
9	ww.hfndawl,com r • e• for »	<1%
10	Submittedto University of Warwick Student Pape	<1%
11	www.Frontlersln.org	<1%
12	S.e.jaal,M. Adam than."Applicationsof Al in Smart Technologiesand Manufacturing", CTIC	•1•

Authorize the system tDedit this file. Authorize

Edft Annotate File8 Sign Convert All

12 S.P. }anI, M. Adam Kha n. "Applications Df Ai in Sman Technologies *and* Manufacturing", CPC Press, 202fi <1%

13 N. V.paganMohan, B.H. V. S. Rama XrishnamRaju, V. Chandra Sekhar, T. V. G. P. Prasad. "Algorithms In Advanced Auirioai Intelligence- Proceedings of International Conferenceon Aigorchms In Advanced Artiflcfai Intelligence (ICAAAI-2024)", CRC Press, 2025 ‹1

14 Arv!nd Dagur, Sortit Agarwal, Dhirendra Xumar"Shukla, Shablr Ali, Sandhya Sharma. "'Arrificlal Jnfelllgence aod Sustainable Innovation - Volume 3", CRC Press, 20*6 ‹1

15 Submitted roTilbury Unlverslry
Student Paper <1%

16 Fi ay KHmm. 5Ongeet y Pini, Krt$hn.a Dec Khmer, Mun ed ju‹n, "Hanrlhock of AI n Engineering Ap¿l‹ca ons - Tools,Techniques. and ATgorithms ', CEC Press, Z025 <1%

17 www.thefreelibrary.com
Internet Source <1%

18 eprints.utar.edu.my
Internet Source <1%

19 Guilin Li, Xing Gao. "The Feature Compressi'on Algorithmsfor Identltying Cytoklner 8ased on CNT Features", IEEE Access, 2Q20 ,n..... <1%

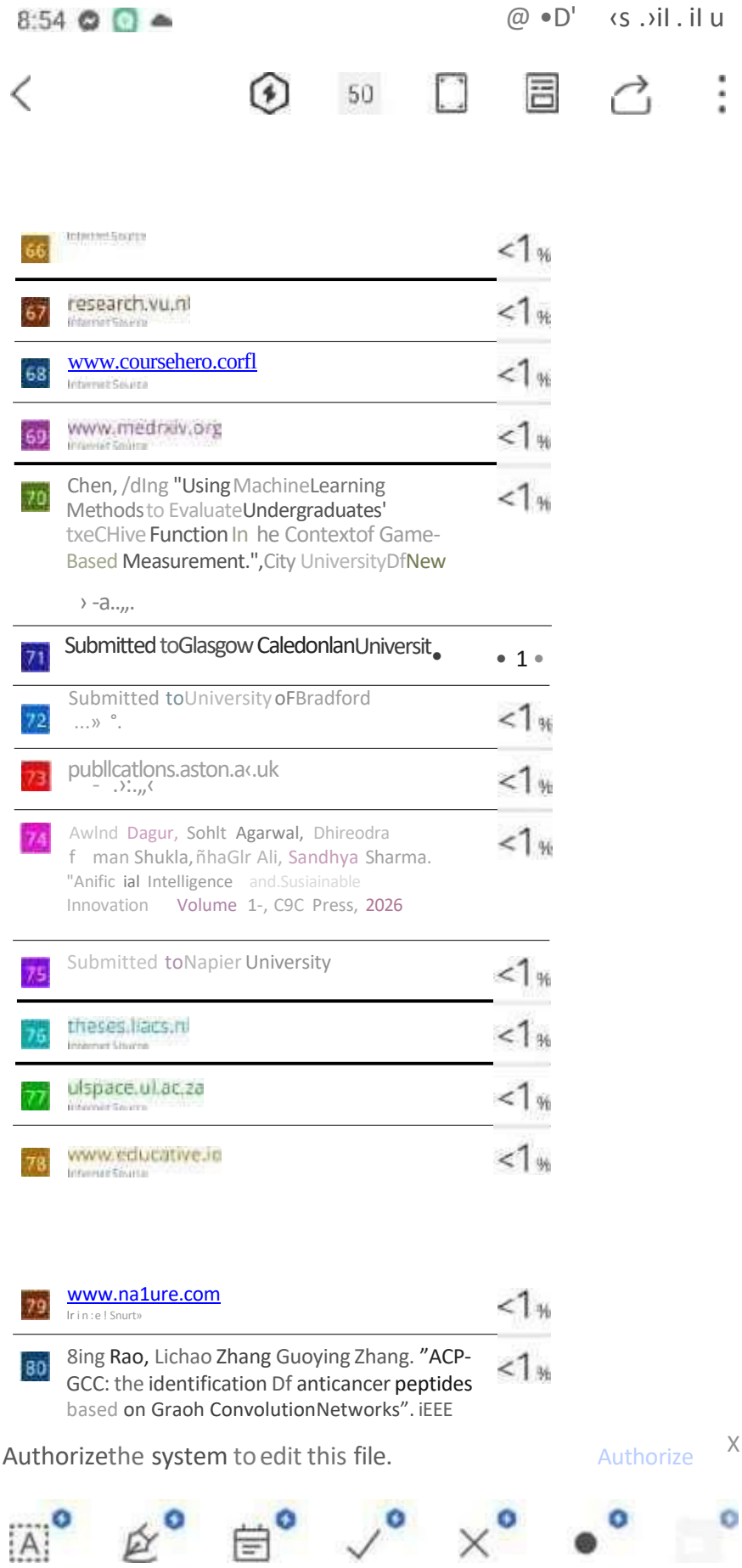
20 Haodong Bian, Maozu Guo,Juan Wang. "Recognition ofMlrochondrialProteins In Plasmodium 8ased on the Tripeptlde Composition", Frorltiérs in Cell and Developmental Biology, 2020 <1%

29	Submitted to University of Birmingham Student Paper	<1%
30	Xuțu M.eng, F-‹Jyi L , Jlnx‹ong then, Junlong Li, Hao Wu, Sfiuqin 11,llangning Song. Çuanzhong L›u "Lurge'sc•ale compar.4r+ve rr:view and assc ssment rif compul.aticn.zl nletriods fur artti-cancer peptide ‹drT1I Íl"Ca|lcw ", UFi e f'nğs 1s BidIn fo r'malics, 2020	<1%
31	esource.dbs.ie	<1%
32	www.preprints,org Internet Source	<1%
33	ÓOEŁOF|g BPI@UÏf1.EO Fit i. i,' -.i':..i.•iii -	<1%
34	umpir.ump.edu.my	<1%
35	www.worldscientific.com .. --.!'t..,..,	<1%
36	Manoj Xurñar, Tanweer Ali, jaume Anguera, Suman Lata Tripafhi. "Emerging Technológiès in AJ, Computarlon, Communication,and CyberseEMFit§ - Proceedlngs *ol* the First International Conference on Artlficlal Intelligence,Computation, Communication and Network Security (A1CCoNS.20ZS)". CRC Press 1026	<1%
37	Mandela Lime, Toßias kłeßmer, Gabriel Micard,Gunnar ScßLibert. "Power grid operation in dlstributlon grids wlth convo!utional neural networks", Smart Ecergy, 2025	<1%
38	Marketa Zveieò‹l. "'understandlng B›oinformatics",Garland Science,2007	‹1

Authorize the system to edit this file. Authorize X

Edit Annotate Fill 8 Sign Čonvert All

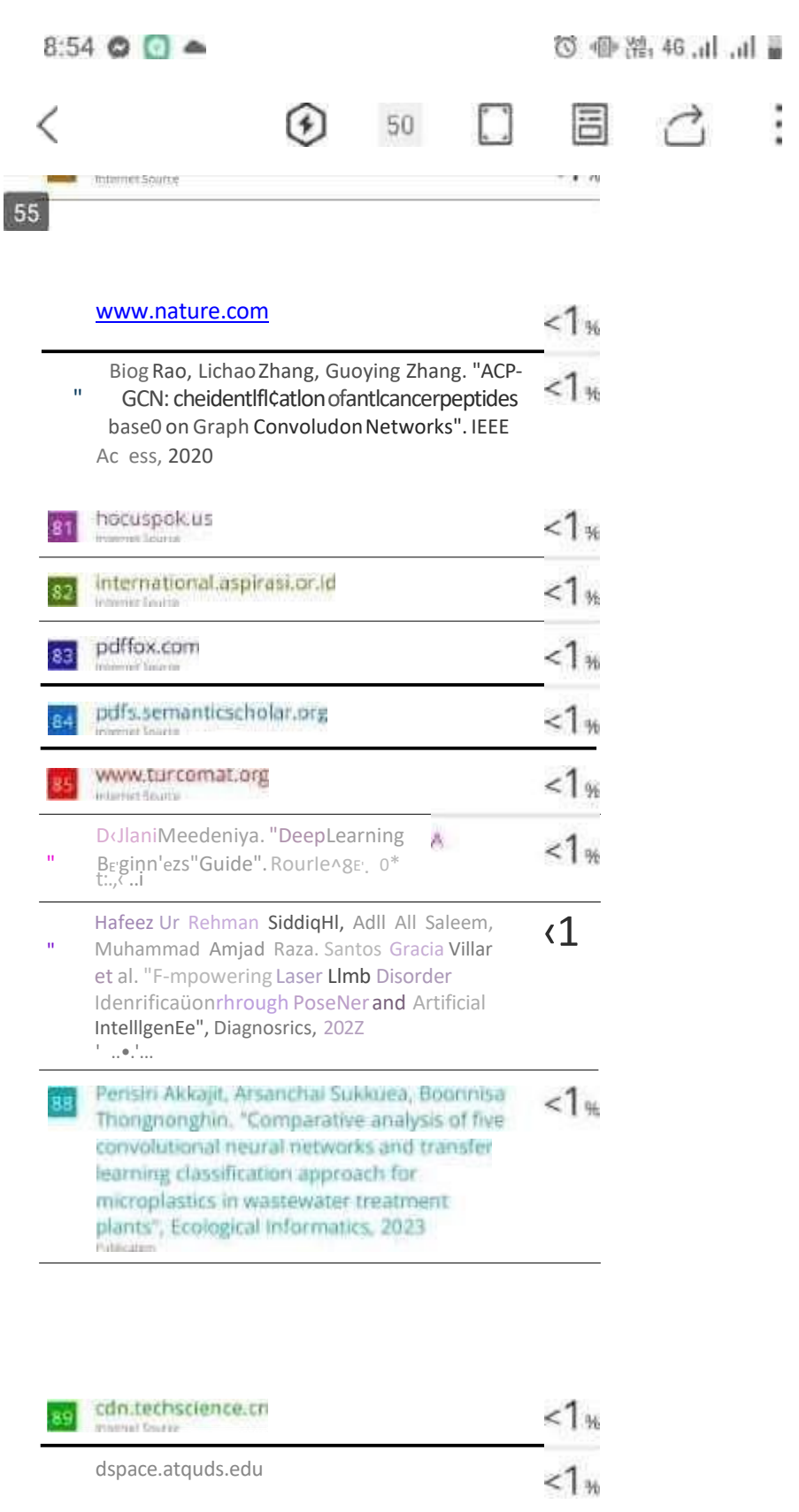
38	Markera Zvelebil.*Unders andlng Bloinfofmat ", Garland Science,2007	<1 %
39	Shantong Hu, Xlaoyu Wang, Zhikang Wang Menghan jJan Shihu| Wang, }iangnfng Song, Gnimfn Zhang. "HPC|as A data-drlven approach for idenlifyirig halophlllc protelns based on cacBoos ", Cold Sprfng Harbor Laboratary, 2023	<1 %
40	fastercapital.com	<1 %
41	oulurepo.oulu.fi	<1 %
42	A. Cfirrril, N. Eeiri,N. Yahyaoul, D.8. Hayraperyan, M.N. Mctrshed."Deep learning-assified prediction of electricfielc! impacton im £/ri\g hOt0lDrli/atlorl Er 5S seCtIOfl In Crl5/Znsecore/shell spheriEa\ quantum dot embedderJIn PVA ma\rir"', Sensors anrJ Acruators A: Physical, 2026	<1 %
43	Oklm Kang,Dayld D.Johnson, Alyssa kermad. "Second LanguageProsody and Computer Modeling", Routledge. 2021	<1 %
44	SukbpreetKaur, Amanpreet Kaur, Manlsb I¢umar. "RecentAdvances in CDmputational Methods in Science andTechnology - Volume CRCPress, 2026	<1 %
45	bmcmedinformdecismak.biomedcenrraLcom	‹1.
52	core.ac.Uk	
46		<1 %
47	lugspace.iogaza.edu,ps	<1 %

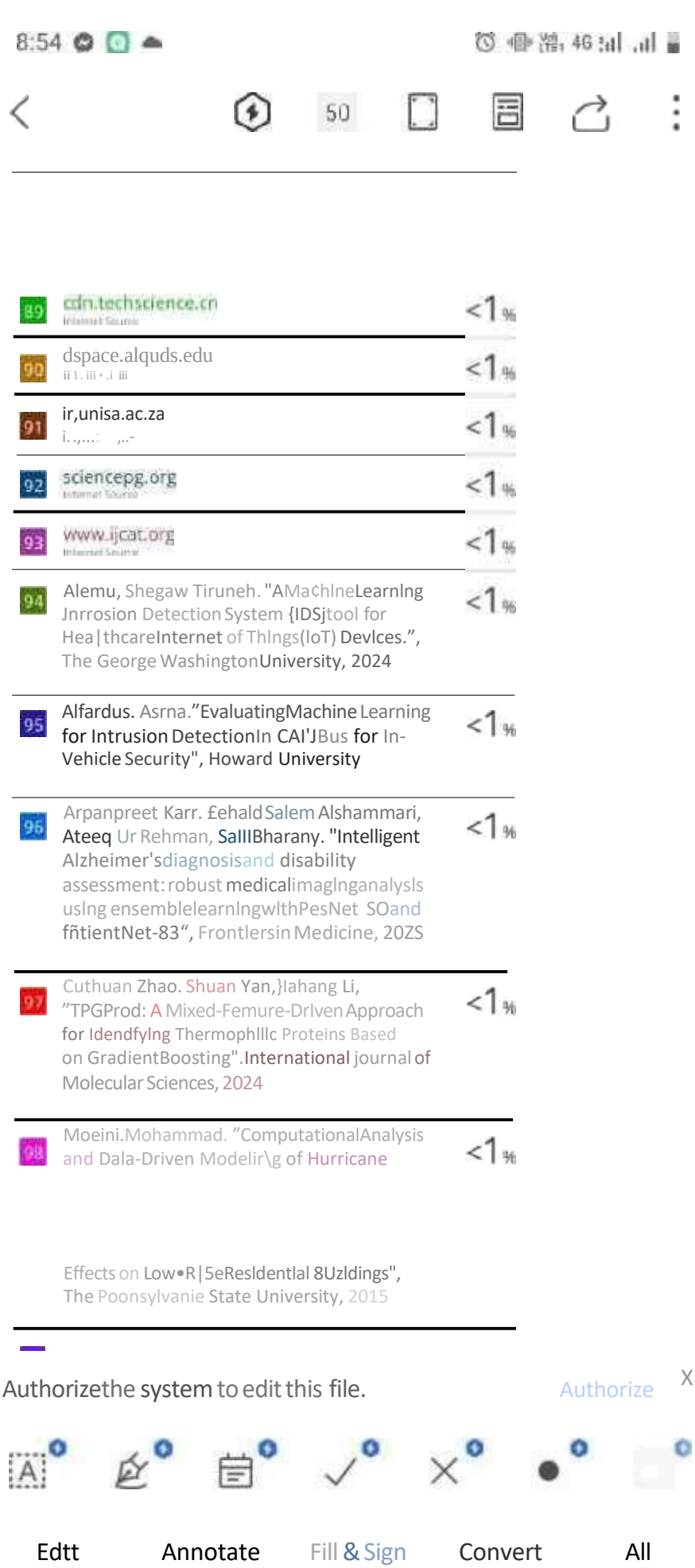


66		<1%
67	research.vu.nl	<1%
68	www.coursehero.corfl	<1%
69	www.medrxiv.org	<1%
70	Chen, /dIng "Using MachineLearning Methods to EvaluateUndergraduates' txeCHive Function In he Contextof Game-Based Measurement.",City UniversityDfNew › -a..,,.	<1%
71	Submitted toGlasgow CaledonlanUniversit.	• 1 •
72	Submitted toUniversity oFBradford ...» °.	<1%
73	publlcatlons.aston.a‹.uk - .›:.„‹	<1%
74	Awlnd Dagur, Sohlt Agarwal, Dhireodra f man Shukla, ñhaGlr Ali, Sandhya Sharma. "Anific ial Intelligence and.Susiainable Innovation Volume 1-, C9C Press, 2026	<1%
75	Submitted toNapier University	<1%
76	theses.liacs.nl	<1%
77	ulspace.ul.ac.za	<1%
78	www.educative.io	<1%
79	www.na1ure.com Ir i n :e ! Snurt»	<1%
80	8ing Rao, Lichao Zhang Guoying Zhang. "ACP-GCC: the identification Df anticancer peptides based on Graoh ConvolutionNetworks". iEEE	<1%

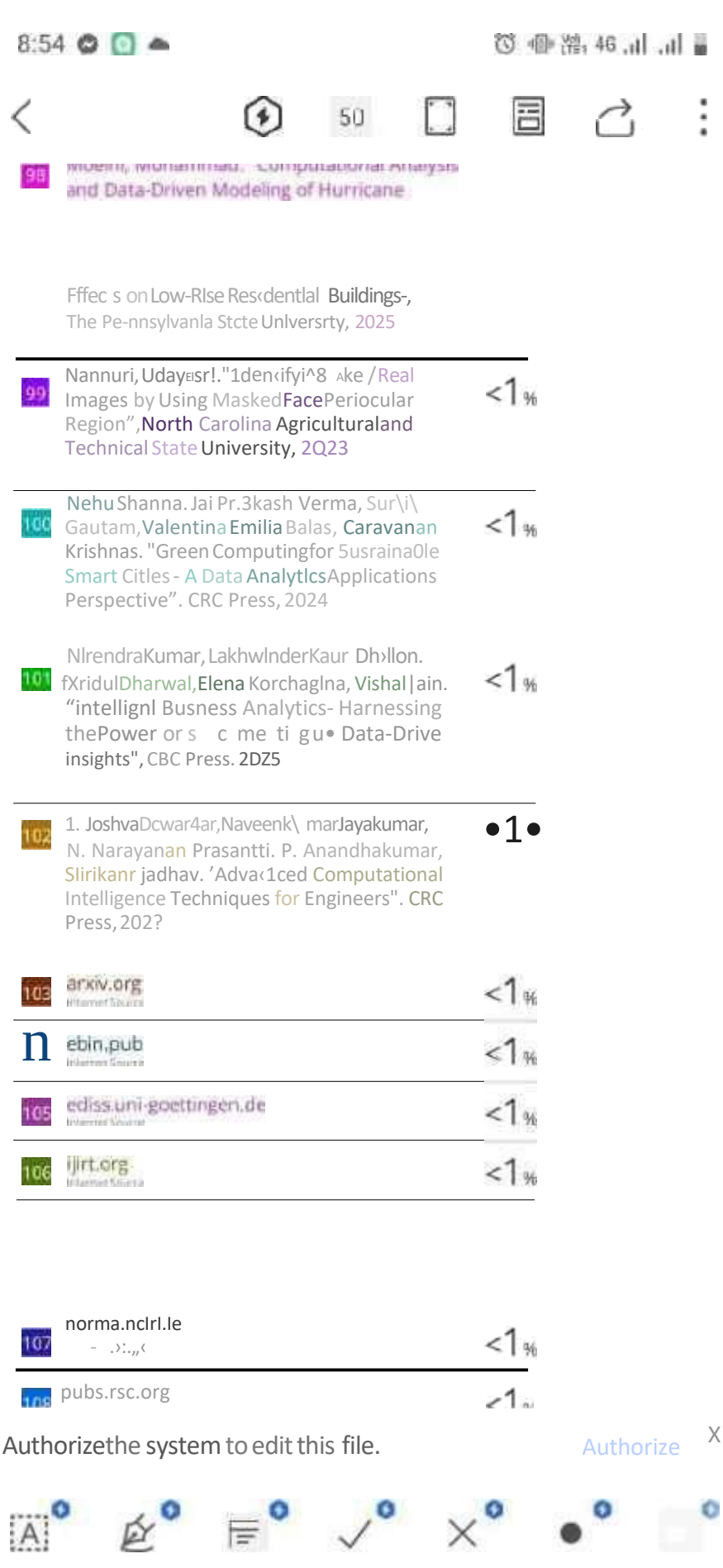
Authorizethe system to edit this file. Authorize X

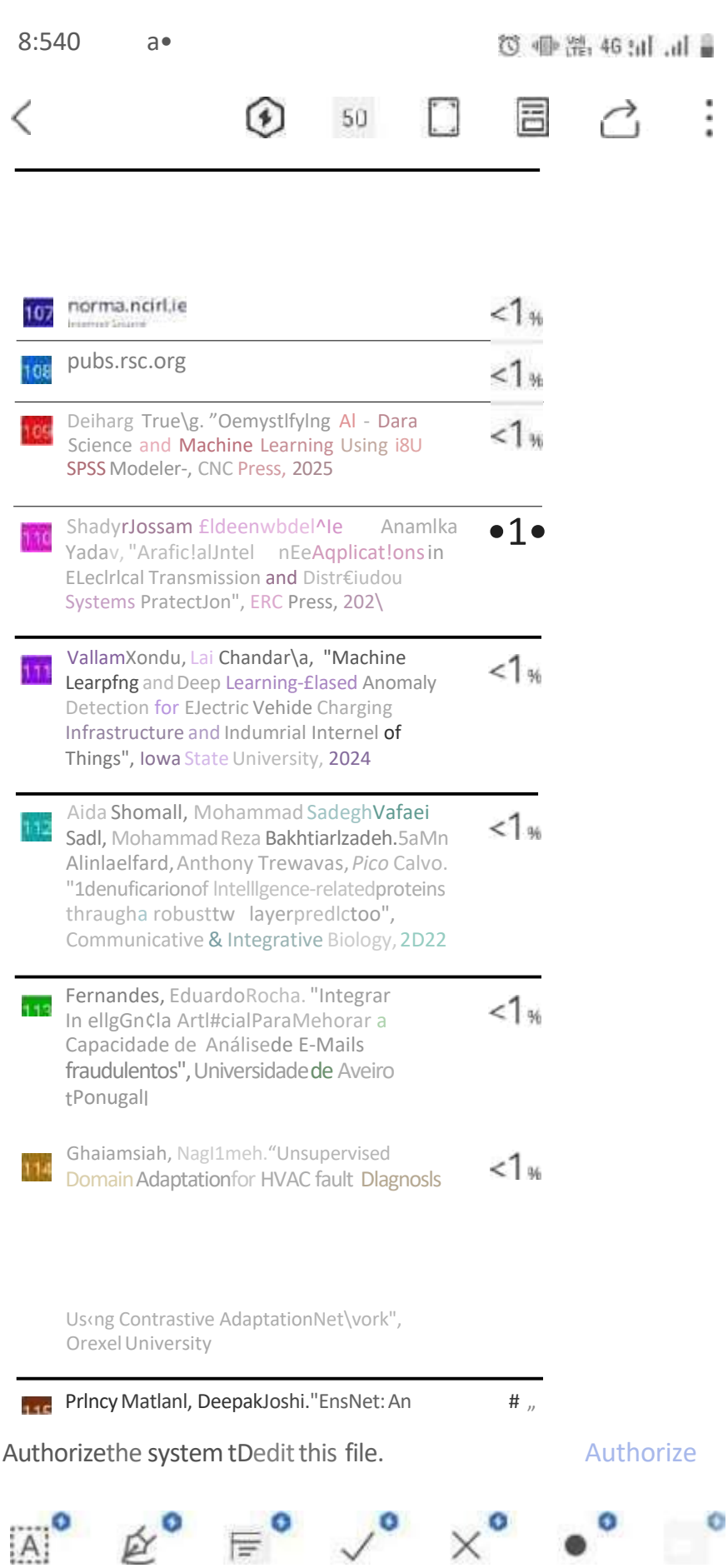
Edtt Annotate Fill & Sign Convert All


8:54
50
55
www.nature.com
<1%
Biog Rao, Lichao Zhang, Guoying Zhang. "ACP-GCN: cheidentlfl¢atlon ofantlcancerpeptides base0 on Graph Convoludon Networks". IEEE Ac ess, 2020
<1%
hocuspok.us
<1%
international.aspirasi.or.id
<1%
pdffox.com
<1%
pdfs.semanticscholar.org
<1%
www.turcomat.org
<1%
D‹JlaniMeedeniya. "DeepLearning Beginn'ezs"Guide". Rourle^gE. 0*
<1%
Hafeez Ur Rehman SiddiqHl, Adll All Saleem, Muhammad Amjad Raza. Santos Gracia Villar et al. "F-mpowering Laser Llmb Disorder Idenrificaüonrhrough PoseNer and Artificial IntelllgenEe", Diagnosrics, 202Z
‹1
Perisiri Akkajit, Arsanchai Sukkuea, Boonnisa Thongnonghin. "Comparative analysis of five convolutional neural networks and transfer learning classification approach for microplastics in wastewater treatment plants", Ecological Informatics, 2023
<1%
cdn.techscience.cn
<1%
dspace.atquds.edu
<1%
Authorize the system to edit this file.
Authorize
X
Edit
Annotate
Fill & Sign
Convert
All



89	cdn.techscience.cn	<1%
90	dspace.alquds.edu	<1%
91	ir,unisa.ac.za	<1%
92	sciencepg.org	<1%
93	www.ijcat.org	<1%
94	Alemu, Shegaw Tiruneh. "AMa¢hlneLearnlng Jnrrosion Detection System {IDSjtool for Hea|thcareInternet of Thlngs(loT) Devlces.", The George WashingtonUniversity, 2024	<1%
95	Alfardus. Asrna."EvaluatingMachine Learning for Intrusion DetectionIn CAI'JBus for In-Vehicle Security", Howard University	<1%
96	Arpanpreet Karr. £ehald Salem Alshammari, Ateeq Ur Rehman, SalllBharany. "Intelligent Alzheimer'sdiagnosisand disability assessment: robust medicalimaglnganalysls uslng ensemblelearnlngwlthPesNet SOand fñtientNet-83", Frontlersin Medicine, 20ZS	<1%
97	Cuthuan Zhao. Shuan Yan,}lahang Li, "TPGProd: A Mixed-Femure-Drlven Approach for Idendfylng Thermophlllc Proteins Based on GradientBoosting".International journal of Molecular Sciences, 2024	<1%
98	Moeini.Mohammad. "ComputationalAnalysis and Dala-Driven Modelir\g of Hurricane Effects on Low•R|5eResldentlal 8Uzldings", The Poonsylvanie State University, 2015	<1%


8:54
50
and Data-Driven Modeling of Hurricane
Fffec s on Low-RIse Res‹dentlal Buildings-,
The Pe-nnsylvanla Stcte Unlversrty, 2025
Nannuri, Uday sr!."1den‹ifyi^8 ake /Real Images by Using MaskedFacePeriocular Region",North Carolina Agriculturaland Technical State University, 2Q23
<1%
Nehu Shanna. Jai Pr.3kash Verma, Sur\i\ Gautam,Valentina Emilia Balas, Caravanan Krishnas. "Green Computingfor 5usraina0le Smart Citles - A Data AnalytlcsApplications Perspective". CRC Press, 2024
<1%
NlrendraKumar, LakhwlnderKaur Dh›llon. fXridulDharwal,Elena Korchaglna, Vishal|ain. "intellignl Busness Analytics- Harnessing thePower or s c me ti gu• Data-Drive insights", CBC Press. 2DZ5
<1%
1. JoshvaDcwar4ar,Naveenk\ marJayakumar, N. Narayanan Prasantti. P. Anandhakumar, Slirikanr jadhav. 'Adva‹1ced Computational Intelligence Techniques for Engineers". CRC Press, 202?
•1•
arxiv.org
<1%
ebin.pub
<1%
ediss.uni-goettingen.de
<1%
ijirt.org
<1%
norma.nclrl.le
<1%
pubs.rsc.org
Authorizethe system to edit this file.
Authorize
X
Edit
Annotate
Fill & Sign
Convert
All



107	norma.ncirl.ie	<1%
108	pubs.rsc.org	<1%
109	Deiharg True\g. ”Oemystlfylng Al - Dara Science and Machine Learning Using i8U SPSS Modeler-, CNC Press, 2025	<1%
110	ShadyrJossam £ldeenwbdel^Ie Anamlka Yadav, "Arafic!alJntel nEeAqplicat!ons in ELeclrlcal Transmission and Distr€iudou Systems PratectJon", ERC Press, 202\	•1•
111	VallamXondu, Lai Chandar\a, "Machine Learpfng and Deep Learning-£lased Anomaly Detection for EJectric Vehide Charging Infrastructure and Indumrial Internel of Things", Iowa State University, 2024	<1%
112	Aida Shomall, Mohammad SadeghVafaei Sadl, Mohammad Reza Bakhtiarlzadeh.5aMn Alinlaelfard, Anthony Trewavas, *Pico* Calvo. "1denuficarionof lntelllgence-relatedproteins thraugha robusttw layerpredlctoo", Communicative & Integrative Biology, 2D22	<1%
113	Fernandes, EduardoRocha. "Integrar In ellgGn¢la Artl#cialParaMehorar a Capacidade de Análisede E-Mails fraudulentos", Universidade de Aveiro tPonugall	<1%
114	Ghaiamsiah, Nagl1meh.“Unsupervised Domain Adaptationfor HVAC fault Dlagnosls	<1%
	Us‹ng Contrastive AdaptationNet\vork", Orexel University	
115	Prlncy Matlanl, DeepakJoshi."EnsNet: An	# „

Authorizethe system tDedit this file. Authorize

Edft Annotate Fill 8 Sign Convert All

Dashboard

Total Payable

747,200.00

Total Paid

747,001.00

Total Due

199.00

Total Other

4,100.00

Transactions History | Date Wise Transactions

Filter By Service:

Choose Any Service