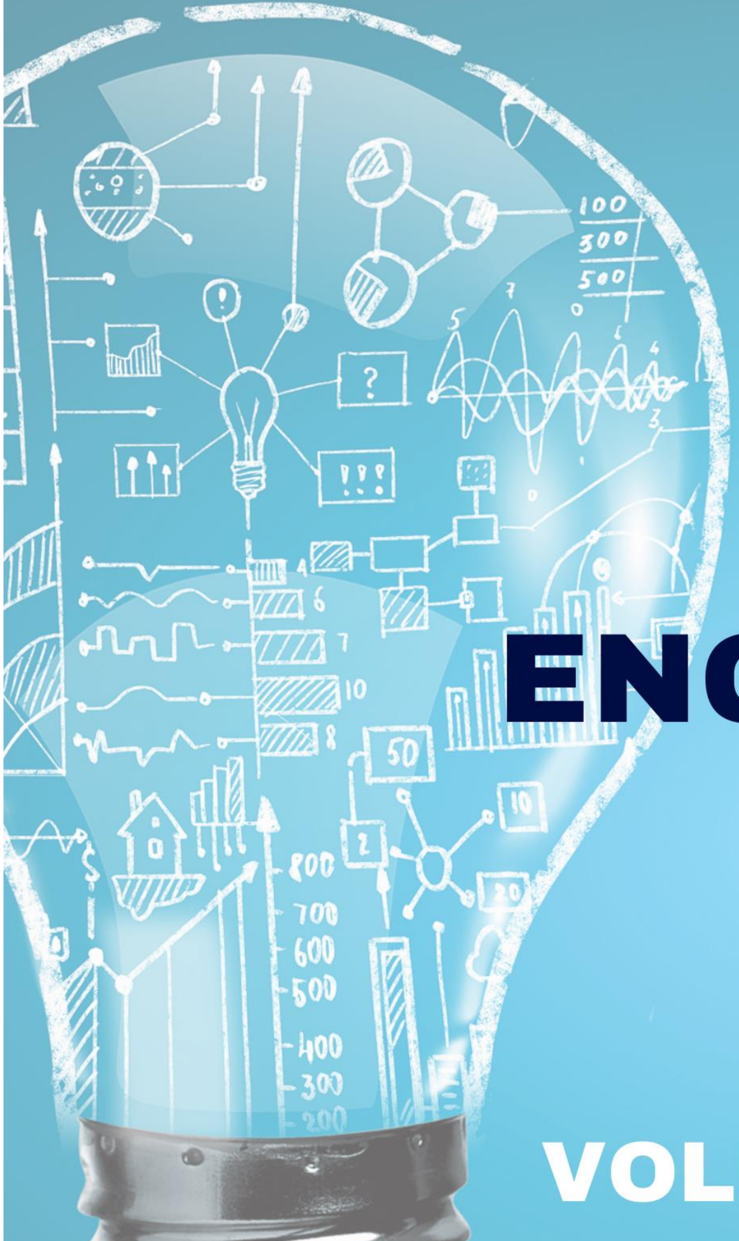
IJOER
ENGINEERING JOURNAL



ENGINEERING JOURNAL IJOER

VOLUME-10, ISSUE-7
JULY 2024

DOWNLOAD

Preface

We would like to present, with great pleasure, the inaugural volume-10, Issue-7, July 2024, of a scholarly journal, *International Journal of Engineering Research & Science*. This journal is part of the AD Publications series *in the field of Engineering, Mathematics, Physics, Chemistry and science Research Development*, and is devoted to the gamut of Engineering and Science issues, from theoretical aspects to application-dependent studies and the validation of emerging technologies.

This journal was envisioned and founded to represent the growing needs of Engineering and Science as an emerging and increasingly vital field, now widely recognized as an integral part of scientific and technical investigations. Its mission is to become a voice of the Engineering and Science community, addressing researchers and practitioners in below areas:

Chemical Engineering	
Biomolecular Engineering	Materials Engineering
Molecular Engineering	Process Engineering
Corrosion Engineering	
Civil Engineering	
Environmental Engineering	Geotechnical Engineering
Structural Engineering	Mining Engineering
Transport Engineering	Water resources Engineering
Electrical Engineering	
Power System Engineering	Optical Engineering
Mechanical Engineering	
Acoustical Engineering	Manufacturing Engineering
Optomechanical Engineering	Thermal Engineering
Power plant Engineering	Energy Engineering
Sports Engineering	Vehicle Engineering
Software Engineering	
Computer-aided Engineering	Cryptographic Engineering
Teletraffic Engineering	Web Engineering
System Engineering	
Mathematics	
Arithmetic	Algebra
Number theory	Field theory and polynomials
Analysis	Combinatorics
Geometry and topology	Topology
Probability and Statistics	Computational Science
Physical Science	Operational Research
Physics	
Nuclear and particle physics	Atomic, molecular, and optical physics
Condensed matter physics	Astrophysics
Applied Physics	Modern physics
Philosophy	Core theories

Chemistry	
Analytical chemistry	Biochemistry
Inorganic chemistry	Materials chemistry
Neurochemistry	Nuclear chemistry
Organic chemistry	Physical chemistry
Other Engineering Areas	
Aerospace Engineering	Agricultural Engineering
Applied Engineering	Biomedical Engineering
Biological Engineering	Building services Engineering
Energy Engineering	Railway Engineering
Industrial Engineering	Mechatronics Engineering
Management Engineering	Military Engineering
Petroleum Engineering	Nuclear Engineering
Textile Engineering	Nano Engineering
Algorithm and Computational Complexity	Artificial Intelligence
Electronics & Communication Engineering	Image Processing
Information Retrieval	Low Power VLSI Design
Neural Networks	Plastic Engineering

Each article in this issue provides an example of a concrete industrial application or a case study of the presented methodology to amplify the impact of the contribution. We are very thankful to everybody within that community who supported the idea of creating a new Research with IJOER. We are certain that this issue will be followed by many others, reporting new developments in the Engineering and Science field. This issue would not have been possible without the great support of the Reviewer, Editorial Board members and also with our Advisory Board Members, and we would like to express our sincere thanks to all of them. We would also like to express our gratitude to the editorial staff of AD Publications, who supported us at every stage of the project. It is our hope that this fine collection of articles will be a valuable resource for *IJOER* readers and will stimulate further research into the vibrant area of Engineering and Science Research.

Mukesh Arora

(Chief Editor)

Board Members

Dr. M. Varatha Vijayan

Annauniversity Rank Holder, Commissioned Officer Indian Navy, Ncc Navy Officer (Ex-Serviceman Navy), Best Researcher Awardee, Best Publication Awardee, Tamilnadu Best Innovation & Social Service Awardee From Lions Club.

Dr. Mohamed Abdel Fatah Ashabrawy Moustafa

PhD. in Computer Science - Faculty of Science - Suez Canal University University, 2010, Egypt.

Assistant Professor Computer Science, Prince Sattam bin AbdulAziz University ALkharj, KSA.

Prof.S.Balamurugan

Dr S. Balamurugan is the Head of Research and Development, Quants IS & CS, India. He has authored/co-authored 35 books, 200+ publications in various international journals and conferences and 6 patents to his credit. He was awarded with Three Post-Doctoral Degrees - Doctor of Science (D.Sc.) degree and Two Doctor of Letters (D.Litt) degrees for his significant contribution to research and development in Engineering.

Dr. Mahdi Hosseini

Dr. Mahdi did his Pre-University (12th) in Mathematical Science. Later he received his Bachelor of Engineering with Distinction in Civil Engineering and later he Received both M.Tech. and Ph.D. Degree in Structural Engineering with Grade "A" First Class with Distinction.

Dr. Anil Lamba

Practice Head – Cyber Security, EXL Services Inc., New Jersey USA.

Dr. Anil Lamba is a researcher, an innovator, and an influencer with proven success in spearheading Strategic Information Security Initiatives and Large-scale IT Infrastructure projects across industry verticals. He has helped bring about a profound shift in cybersecurity defense. Throughout his career, he has parlayed his extensive background in security and a deep knowledge to help organizations build and implement strategic cybersecurity solutions. His published researches and conference papers has led to many thought provoking examples for augmenting better security.

Dr. Ali İhsan KAYA

Currently working as Associate Professor in Mehmet Akif Ersoy University, Turkey.

Research Area: Civil Engineering - Building Material - Insulation Materials Applications, Chemistry - Physical Chemistry – Composites.

Dr. Parsa Heydarpour

Ph.D. in Structural Engineering from George Washington University (Jan 2018), GPA=4.00.

Dr. Heba Mahmoud Mohamed Afify

Ph.D degree of philosophy in Biomedical Engineering, Cairo University, Egypt worked as Assistant Professor at MTI University.

Dr. Dibya Prakash Rai

Post Doctoral Fellow (PDF), M.Sc., B.Sc., Working as Assistant Professor in Department of Physics in Pachhuncga University College, Mizoram, India.

Dr. Pankaj Kumar Pal

Ph.D R/S, ECE Deptt., IIT-Roorkee.

Dr. P. Thangam

PhD in Information & Communication Engineering, ME (CSE), BE (Computer Hardware & Software), currently serving as Associate Professor in the Department of Computer Science and Engineering of Coimbatore Institute of Engineering and Technology.

Dr. Pradeep K. Sharma

PhD., M.Phil, M.Sc, B.Sc, in Physics, MBA in System Management, Presently working as Provost and Associate Professor & Head of Department for Physics in University of Engineering & Management, Jaipur.

Dr. R. Devi Priya

Ph.D (CSE), Anna University Chennai in 2013, M.E, B.E (CSE) from Kongu Engineering College, currently working in the Department of Computer Science and Engineering in Kongu Engineering College, Tamil Nadu, India.

Dr. Sandeep

Post-doctoral fellow, Principal Investigator, Young Scientist Scheme Project (DST-SERB), Department of Physics, Mizoram University, Aizawl Mizoram, India- 796001.

Dr. Roberto Volpe

Faculty of Engineering and Architecture, Università degli Studi di Enna "Kore", Cittadella Universitaria, 94100 – Enna (IT).

Dr. S. Kannadhasan

Ph.D (Smart Antennas), M.E (Communication Systems), M.B.A (Human Resources).

Research Area: Engineering Physics, Electromagnetic Field Theory, Electronic Material and Processes, Wireless Communications.

Mr. Amit Kumar

Amit Kumar is associated as a Researcher with the Department of Computer Science, College of Information Science and Technology, Nanjing Forestry University, Nanjing, China since 2009. He is working as a State Representative (HP), Spoken Tutorial Project, IIT Bombay promoting and integrating ICT in Literacy through Free and Open Source Software under National Mission on Education through ICT (NMEICT) of MHRD, Govt. of India; in the state of Himachal Pradesh, India.

Mr. Tanvir Singh

Tanvir Singh is acting as Outreach Officer (Punjab and J&K) for MHRD Govt. of India Project: Spoken Tutorial - IIT Bombay fostering IT Literacy through Open Source Technology under National Mission on Education through ICT (NMEICT). He is also acting as Research Associate since 2010 with Nanjing Forestry University, Nanjing, Jiangsu, China in the field of Social and Environmental Sustainability.

Mr. Abilash

MTech in VLSI, BTech in Electronics & Telecommunication engineering through A.M.I.E.T.E from Central Electronics Engineering Research Institute (C.E.E.R.I) Pilani, Industrial Electronics from ATI-EPI Hyderabad, IEEE course in Mechatronics, CSHAM from Birla Institute Of Professional Studies.

Mr. Varun Shukla

M.Tech in ECE from RGPV (Awarded with silver Medal By President of India), Assistant Professor, Dept. of ECE, PSIT, Kanpur.

Mr. Shrikant Harle

Presently working as a Assistant Professor in Civil Engineering field of Prof. Ram Meghe College of Engineering and Management, Amravati. He was Senior Design Engineer (Larsen & Toubro Limited, India).

Zairi Ismael Rizman

Senior Lecturer, Faculty of Electrical Engineering, Universiti Teknologi MARA (UiTM) (Terengganu) Malaysia Master (Science) in Microelectronics (2005), Universiti Kebangsaan Malaysia (UKM), Malaysia. Bachelor (Hons.) and Diploma in Electrical Engineering (Communication) (2002), UiTM Shah Alam, Malaysia.

Mr. Ronak

Qualification: M.Tech. in Mechanical Engineering (CAD/CAM), B.E.

Presently working as a Assistant Professor in Mechanical Engineering in ITM Vocational University, Vadodara. Mr. Ronak also worked as Design Engineer at Finstern Engineering Private Limited, Makarpura, Vadodara.

Table of Contents

Volume-10, Issue-7, July 2024

GILPI: Graphlet Interaction - based lncRNA-Protein Interaction Prediction

Hong-Yi Zhang[1*], Yan Zhou[2]

[*1]Big data statistics specialization, Guizhou University of Finance and Economics, China-GuiYang
[2]Specialization in subject teaching (language), Guizhou Normal University, China-GuiYang
*Corresponding Author

Received: 01 July 2024/ Revised: 08 July 2024/ Accepted: 15 July 2024/ Published: 31-07-2024

***Abstract*— *Identification of lncRNA-protein interactions is important for understanding the biological functions and molecular mechanisms of lncRNAs. In this study, we proposed a computational model for predicting lncRNA-protein interactions based on Graphlet interactions to find potential LPIs (GILPI). First, five LPI datasets were collected. Second, vector features of lncRNAs and proteins were extracted from the sequence data by pyfeat and BioTriangle, respectively. Third, these features were subjected to Pearson's correlation coefficient to calculate the similarity between lncRNAs and the similarity between proteins. Fourth, the Jaccard similarity between lncRNAs and proteins was calculated based on the LPI network, and then the corresponding Pearson similarity and Jaccard similarity were taken as the average value of the final lncRNA-lncRNA similarity and protein-protein similarity to construct the network. Finally, lncRNA-protein classification prediction was performed on both networks. Comparing GILPI with five state-of-the-art LPI prediction methods through 5-fold cross-validation, the results show that the GILPI prediction model has strong LPI classification performance. The case studies show that there may be interactions between NONHSAT021830 and Q9H9S0, n385685 and Q07955, and NONHSAT098243 and P25490.The novelty of GILPI is that it integrates the two similarities to construct a network, and then utilizes Graphlet interactions on the network to directly and indirectly link the features to mine out potential features, thus greatly improving the performance of the model.*

***Keywords*— Graphlet interaction, Jaccard similarity, Pearson similarity, lncRNA-protein interaction.**

I. INTRODUCTION

1.1 Motivation:

Long non-coding RNAs (lncRNAs) are transcripts composed of more than 200 nucleotides but lack coding capabilities [1]. lncRNAs play key roles in biological processes such as gene expression regulation, epigenetic regulation, and cell differentiation [2].

For example, HOXA-AS2 and SNHG12 in lncRNAs have been identified as potential therapeutic targets and biomarkers for human cancers [3]. DLEU1 is closely related to colorectal cancer through activation of KPNA3, the expression of HOTAIR is elevated in lung cancer, and ZFAS1 is closely related to the chemosensitivity of cervical cancer cells [4]. In summary, more and more experiments have confirmed that lncRNAs are tumor-related biomolecules. However, to date, the relationship between lncRNAs and known tumor suppressor entities remains largely elusive. There is evidence that lncRNAs exert their biological functions through binding to RNA-binding proteins. Therefore, identifying potential lncRNA-protein interactions (LPIs) contributes to understanding many important biological processes and the treatment of various complex diseases.

1.2 Related Work:

Identifying lncRNA-protein interactions (LPIs) generally adopts two methods: experimental methods and computational methods. In experimental methods, biologists initially detect lncRNA-protein interactions through bioexperiments, such as RNA pulldown [5], RNA binding protein immunoprecipitation (RIP) [6], etc. However, this method is time-consuming and wasteful of resources. Gradually, people explore potential LPIs with computational methods, mainly divided into machine learning-based methods and network-based methods.

Machine learning-based methods mainly describe lncRNA-protein pairs by selecting features of lncRNAs and proteins, and use the extracted features as input to train a supervised learning model to identify potential LPIs. Liu et al.[7], Zhang et al.[8], Ma et al.[9] explored the neighborhood regularized logistic matrix decomposition method, graph regularized nonnegative matrix factorization model, and projection-based neighborhood nonnegative matrix factorization method (PMKDN), respectively.

Network-based methods usually construct some associated networks of lncRNAs or proteins, and then design a network algorithm to calculate the probability or score of interaction between lncRNAs and proteins. Zhao et al.[10] and Ge et al.[11] designed two recommendation algorithms based on bipartite networks to score each lncRNA-protein pair. Jia[12] et al. proposed a multifeature fusion method based on linear neighborhood propagation to calculate the linear neighborhood similarity of feature space and predict the results through label propagation.

Computational methods can effectively discover many potential relationships between lncRNAs and proteins. However, most machine learning-based LPI prediction methods are measured on a single dataset, which may lead to prediction bias. Secondly, cross-validation is performed in the case of lncRNA-protein pairs, ignoring the performance under other cross-validations. Network-based methods cannot find possible potential associated proteins or lncRNAs for a single lncRNA or protein.

1.3 Research Contributions

In this paper, we developed a network-based LPI prediction model, GILPI, to predict the interaction relationships between lncRNAs and proteins. The GILPI model integrates the bioinformatics of lncRNAs and proteins, Pearson similarity, Jaccard similarity, and Graphlet interactions into a unified prediction framework to identify potential LPIs. The main contributions of this work are as follows:

1) It reasonably integrates a variety of biological characteristics of lncRNAs and proteins, including 13 types for lncRNAs and 14 types for proteins, enabling a more effective description of lncRNA-protein pairs.
2) It creates networks of lncRNAs and proteins composed of Pearson similarity and Jaccard similarity, and utilizes Graphlet interactions on these networks to classify and predict unknown lncRNA-protein pairs.
3) By leveraging the direct and indirect connections of Graphlet interactions, it deeply mines the hidden features between lncRNA-protein pairs, thereby enhancing the predictive performance of GILPI.

II. MATERIALS AND METHODS

2.1 Data Preparation:

2.1.1 Dataset Acquisition:

In this paper, we have compiled five datasets related to LPI. Datasets 1, 2, and 3 contain human LPI data, while Datasets 4 and 5 contain plant LPI data. Dataset 1 is provided by Li et al. [13]. After removing lncRNAs and proteins with unknown sequence information from NPInter[14], NONCODE[15], and UniProt[16], we obtained 3,479 known associations from 935 lncRNAs and 59 proteins. Dataset 2 was constructed by Zheng et al. [17]. After similar preprocessing to Dataset 1, we filtered out 3,265 known associations from 885 lncRNAs and 84 proteins. Dataset 3 was constructed by Zhang et al. [18]. and contains 4,158 interactions from 990 lncRNAs and 27 proteins. Datasets 4 and 5 are from Arabidopsis thaliana and maize, respectively. The former contains 948 interactions from 109 lncRNAs and 35 proteins, while the latter contains 22,133 associations from 1,704 lncRNAs and 42 proteins. The sequence data was extracted from the PlncRNADB database [19], and the interaction data was obtained from http://bis.zju.edu.cn/PlncRNADB/. The five data details are shown in Table 1:

TABLE 1
LPI DATA

Dataset	lncRNAs	Protein	LPIs
Data1	935	59	3479
Data2	885	84	3265
Data3	990	27	4158
Data4	109	35	948
Data5	1704	42	22133

We represent the LPI network as a matrix Y, where elements contain:

$$y(i,j) = \begin{cases} 1, & \text{If lncRNA interacts with protein} \\ 0, & \text{Other} \end{cases} \tag{1}$$

2.1.2 Feature of lncRNAs:

After obtaining the sequence information for the five datasets, we selected 13 features to describe lncRNAs, which are as follows: zCurve, gcContent, atgcRatio, cumulativeSkew, pseudoKNC, monoMonoKGap, mono-DiKGap, monoTriKGap, diMo-noKGap, diDiKGap, diTriK-Gap, triMonoKGap, and tri-DiKGap. The corresponding features were extracted using the Pyfeat [20] Python tool, resulting in a 14,892-dimensional vector.

2.1.3 Feature of Proteins

To describe the biological information of proteins, we selected 14 features, which are as follows: amino acid composition, dipeptide composition, tri-peptide composition, CTD composition, CTD transition, CTD distribution, M-B autocorrelation, Moran autocorrelation, Geary autocorrelation, conjoint triad features, quasi-sequence order descriptors, sequence order coupling number, pseudo amino acid composition 1, and pseudo amino acid composi-tion 2. Features generated by BioTriangle [21] can effectively distinguish the captured amino acid information. In this study, we utilized the BioTriangle software to extract protein features, resulting in a 10,029-dimensional vector.

2.2 Overview of GILPI:

In this study, we created a framework for the LPI prediction model GILPI that integrates Pearson similarity, Jaccard similarity, and Graphlet interactions to classify unknown lncRNA-protein pairs. The following figure describes the GILPI framework.

In Fig. 1, the lncRNA-lncRNA Pearson similarity network, protein-protein Pearson similarity network, lncRNA-lncRNA Jaccard similarity network, protein-protein Jaccard similarity network were obtained after putting the lncRNA vectors and the protein vectors through the Pearson similarity and Jaccard similarity calculations. Then the lncRNA-lncRNA similarity network was constructed by adding the lncRNA-lncRNA Pearson similarity and lncRNA-lncRNA Jaccard similarity and taking the mean value, respectively. The protein-protein similarity network was constructed after summing protein-protein Pearson similarity, protein-protein Jaccard similarity and taking the mean value.

Next, the number of Graphlets is traversed on the lncRNA-lncRNA similarity network and the protein-protein similarity network to train the model. This process yields the weight coefficient V_L for the lncRNA similarity network and the weight coefficient V_P for the protein similarity network. Subsequently, the scores for the test set and the candidate set are calculated to determine the relationships between lncRNAs and proteins.

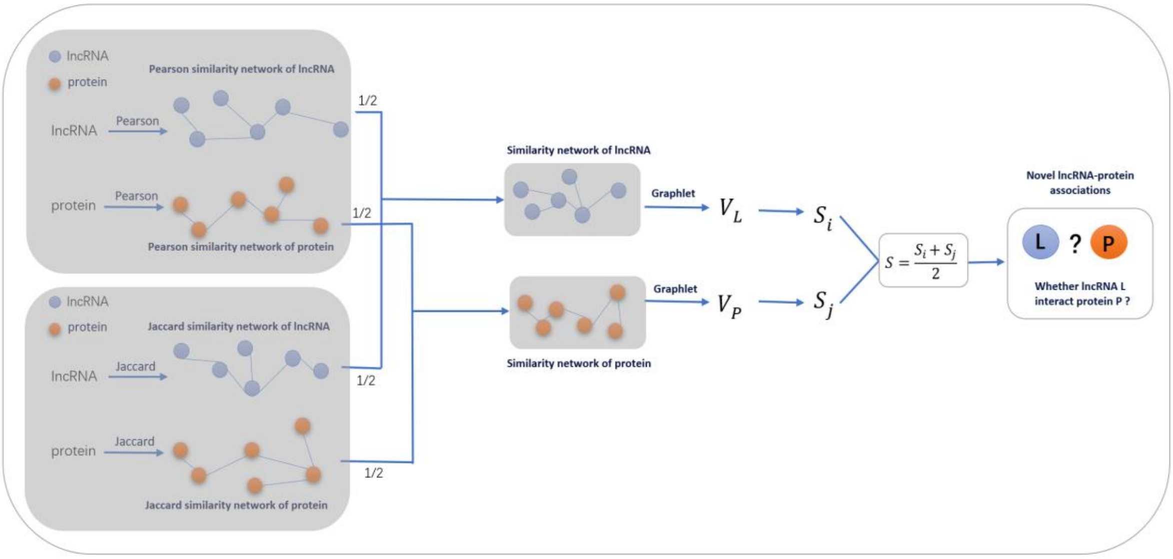

lncRNA
protein
Pearson similarity network of lncRNA
lncRNA
Pearson
protein
Pearson
Pearson similarity network of protein
1/2
1/2
lncRNA
protein
Jaccard similarity network of lncRNA
lncRNA
Jaccard
protein
Jaccard
Jaccard similarity network of protein
1/2
1/2
Similarity network of lncRNA
Graphlet
V_L
S_i
Similarity network of protein
Graphlet
V_P
S_j
$S = \frac{S_i + S_j}{2}$
Novel lncRNA-protein associations
L
?
P
Whether lncRNA L interact protein P ?

FIGURE 1: Framework of GILPI

2.3 Network Construction

2.3.1 Construction of lncRNA-lncRNA Pearson Similarity Network

We used the 14,892-dimensional vectors extracted with the Pyfeat Python tool to calculate the Pearson similarity between lncRNAs using the Pearson correlation coefficient. This resulted in an lncRNA-lncRNA Pearson similarity network. The formula is as follows:

$$\rho_{x,x_1} = \frac{cov(x,x_1)}{\sigma_x \sigma x_1} \tag{2}$$

Where x and $x_1$represent different lncRNAs, $cov(x, x_1)$ is the covariance between two lncRNAs, and $\sigma_x \sigma x_1$ are the standard deviations of the two lncRNAs. The value of ρ_{x,x_1} ranges from -1 to 1, with values less than 0 indicating negative correlation and values greater than 0 indicating positive correlation. The Pearson similarity network for lncRNAs in the five datasets was calculated using this formula.

2.3.2 Protein-protein Pearson similarity network construction:

In order to construct the Pearson similarity network between proteins, we use 10029-dimensional vectors extracted by BioTriangle software and also go through the Pearson correlation coefficient to calculate the Pearson similarity between two proteins two by two to get the protein-protein Pearson similarity network. The formula is the same as shown in (3) above.Just where x and $x_1$denote different proteins respectively, $cov(x, x_1)$ is the covariance between two proteins, andρ_{x,x_1} is the standard deviation between two proteins.

2.3.3 Calculation of lncRNAs and protein Jaccard similarity:

In order to fully explore the biological properties of lncRNAs and proteins, this paper not only used the Pearson correlation coefficient to calculate the similarity between lncRNAs and lncRNAs and between proteins and proteins, but also introduced the Jaccard similarity to measure the relationship between lncRNAs and lncRNAs and between proteins and proteins.

Jaccard similarity is a popular approximation metric for calculating the similarity between two objects. It can be used to find the similarity between two asymmetric binomial vectors or to find the similarity between two sets. Jaccard coefficient is usually used between texts that are sequence order insensitive. The higher the value of Jaccard coefficient, the more similar the samples are. In LPI network, based on known lncRNAs and proteins, we calculated lncRNA-lncRNA similarity and protein-protein similarity by using Jaccard similarity principle. jaccard coefficient is defined as the size of intersection of the sample sets divided by the size of the merged set. For example, L_i and $L_j$are two lncRNA datasets. The Jaccard similarity between any two sets of lncRNAs is calculated as follows:

$$J(L_i, \ L_j) = \frac{|L_i \cap L_j|}{|L_i U L_j|} \tag{3}$$

The Jaccard similarity of proteins was calculated identically to lncRNA.

2.3.4 lncRNAs and protein similarity network construction:

The lncRNA-lncRNA Pearson similarity and protein-protein Pearson similarity were obtained by Pearson correlation calculation, and the lncRNA-lncRNA Jaccard similarity and protein-protein Jaccard similarity were obtained by Jaccard calculation, and then the lncRNA-lncRNA Pearson similarity was added, lncRNA-lncRNA Jaccard similarity and protein-protein Pearson similarity, protein-protein Jaccard similarity were averaged after addition to get the final lncRNA-lncRNA similarity network, protein-protein similarity network.

2.4 Introduction to Graphlet and Graphlet Interaction:

Graphlets are small non-isomorphic connected subgraphs, and a complete large network is composed of Graphlets. In this paper, we only consider Graphlets with no more than 4 nodes, as shown in Fig2 below. In the figure, from G1 to G9 are the 9 types of the corresponding Graphlet, the nodes in the Graphlet occupy different positions called self-isomorphic orbits, and the nodes on the same self-isomorphic orbit have the same local topological features in the Graphlet, and there are 15 self-isomorphic orbits for these 9 types of Graphlets.

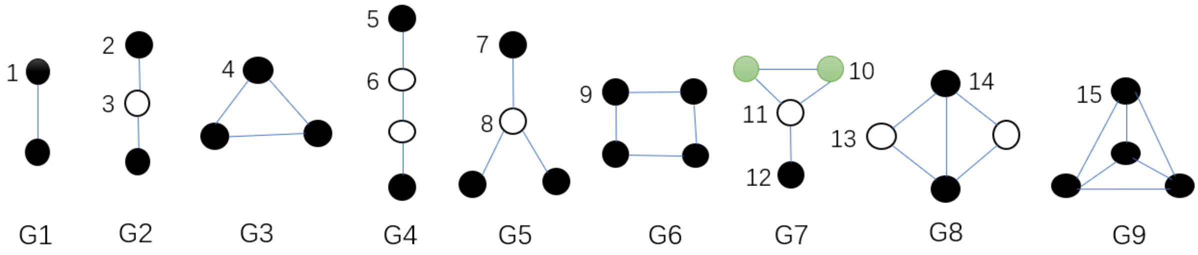


FIGURE 2: Graphlet diagram

Graphlet interaction describes the relationship between 2 nodes. There is a Graphlet interaction between two nodes in the same Graphlet, when there is a Graphlet interaction between node i and node j of graph H, the following equation is satisfied:

$$\exists G \subseteq H,\quad and\ \mathrm{i} \in G,\quad j \in G \tag{4}$$

Where G is a Graphlet in the graph H and V (G) is the set of nodes of G.

In Fig 3 below, the black and light green nodes represent nodes i and j with Graphlet interactions. therefore, different types of relationships exist between two nodes based on their different self-isomorphic orbits. Different types of relationships between two nodes are called Graphlet interaction isomers. For example, Graphlet interaction isomers I_2、$I_3$and I_4. nodes i and j are in different self-isomorphic orbits and are viewed as different Graphlet interaction isomers. graphlet interactions are a vector, where each element denotes the number of corresponding Graphlet interaction isomers. graphlet interaction vector has 28 elements corresponding to 28 Graphlet interaction isomers. In this paper, only Graphlet interactions with no more than 4 nodes are considered, and there are a total of 28 Graphlet interaction isomers, labeled $I_1$to I_{28}.

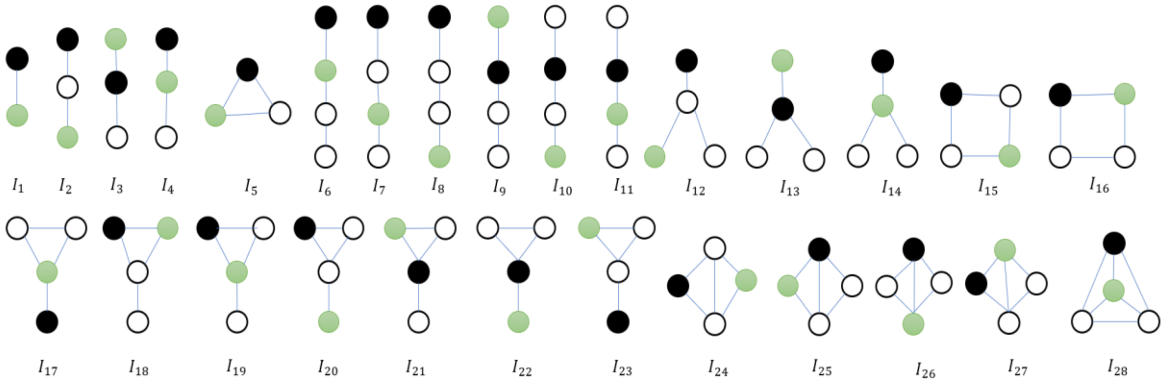


FIGURE 3: Graphlet interactions

2.5 Graphlet interaction computation:

Denote the graph H by the adjacency matrix A= （a_{ij}） . in the graph, a_{ij}=1 if there is an edge between node i and node j, and a_{ij}=0 if there is no edge connecting node i and node j. In the calculation of Graphlet interactions between nodes i and j, the number of isomers I_k is calculated as follows in Eq:

$$N_{ij}(I_k) = \sum_{l\in V(G)} \sum_{m\in V(G)} b_{ij} b_{il} b_{im} b_{jl} b_{jm} b_{lm} \tag{5}$$

To make the above equation clearer, b is a variable and is calculated as follows:

$$b(i,j) = \begin{cases} a_{st}, & \text{s and t are linked in } I_k \\ 1 - a_{st}, & \text{s and t are not linked in } I_k \end{cases} \tag{6}$$

In the above equation, $N_{ij}(I_k)$denotes the number of isomers $I_k$between nodes i and j. l and m denote the other 2 nodes other than nodes i and j. i, j, l, and m are unequal. The above equation calculates the total number of isomers from node i to j. The higher the number of isomers I_k, the closer the relationship between node i and node j is indicated.

Since formula (5) is time consuming to compute, plus in practice isomers are computed from vectors of adjacency matrices such as a_i and a_j, so formula (5) can be rewritten as:

$$N_{ij}(I_k) = a_i * a_j \tag{7}$$

where $*$ denotes the inner product of the two vectors a_i and a_j.

Graphlet has directionality. When calculating the Graphlet interactions of two nodes i and j, the Graphlet interactions from node i to node j are not equal to the Graphlet interactions from node j to node i. The Graphlets have symmetry. However, Graphlets have symmetry, such asI_3 and I_4, I_{17} and I_{22} in Fig3. where $N_{ij}(I_3)$=$N_{ij}(I_4)$ means that the 3rd element of the Graphlet interaction vector from node i to node j is equal to the 4th element of the Graphlet interaction vector from node j to node i.

2.6 Sorting LPIs for unknown associations based on Graphlet interaction scores

Graphlet interactions were used to categorize LPIs, and PLIs with unknown associations were sorted based on Graphlet interaction scores. The higher the score, the more closely the LPI in which the lncRNA is likely to be related to the protein. Below is the formula for calculating the Graphlet interaction score in the protein-protein similarity network:

$$S_j = \sum_k v_k \sum_{i\in P} norm\left(N_{ij}(I_k)\right) \tag{8}$$

In the above equation, S_j denotes the protein-to-protein fraction in the network, P denotes the set of points with known associations for a particular class of proteins, v_k denotes the corresponding weight coefficients, and $norm\left(N_{ij}(I_k)\right)$ denotes the Graphlet interactions normalized from node i to node j. The normalization formula is as follows:

$$norm\left(N_{ij}(I_k)\right) = \frac{N_{ij}(I_k)}{N_i(I_k)} \tag{9}$$

Where $N_{ij}(I_k)$ denotes the number of Graphlet interaction isomersI_k between node i to node j, and $N_i(I_k)$ is the total number of Graphlet interaction isomers I_k between node i to all other nodes. $N_i(I_k)$ is calculated as follows:

$$N_i(I_k) = \sum_{j\in C} N_{ij}(I_k) \tag{10}$$

Where C denotes the set of unknown associations of a certain protein.

The weight v_k in Eq. (8), we use linear regression to calculate. In order to validate the performance of the proposed algorithmic model in this paper, we divide the data into training set and test set. The training set is put through regression to get the weights, and then the test set is used to validate the algorithmic model.

Rewrite equation (8) as:

$$S_j = \sum_k v_k x_{jk} \tag{11}$$

where x_{jk} is calculated by the following equation:

$$x_{jk} = \sum_{i \in P} norm\left(N_{ij}(I_k)\right) \tag{12}$$

At the time of training data, S_j and x_{jk} in Eq. (11) are known, so it is possible to calculate v_k, which is given below:

$$V = (XX^T)^{-1}XS \tag{13}$$

Similarly, in the lncRNA network, the lncRNA-to-lncRNA Graphlet interaction score is given by:

$$S_i = \sum_k v_k \sum_{j \in R} norm\left(N_{ij}(I_k)\right) \tag{14}$$

In Eq. S_i denotes the lncRNAs to lncRNAs score, R denotes the set of points with known associations for a particular type of lncRNA, v_k denotes the corresponding weight coefficients, and $norm\left(N_{ij}(I_k)\right)$ denotes the node i to node j normalized Graphlet interaction. Similarly, Eq. (14) can be rewritten as:

$$S_i = \sum_k v_k x_{ik} \tag{15}$$

The following equation calculates x_{ik}:

$$x_{ik} = \sum_{j \in R} norm\left(N_{ij}(I_k)\right) \tag{16}$$

Finally, the protein-to-protein fraction S_j and the lncRNA-to-lncRNA fraction S_i were used to take the mean value as the calculated protein-to-lncRNA fraction S, calculated as follows:

$$S = \frac{S_i + S_j}{2} \tag{17}$$

III. RESULTS

3.1 Performance Evaluation

To evaluate the performance of the GILPI model, we use five-fold cross-validation to rank the test samples and candidate samples on each of the five datasets, calculate the AUC and AUPR, and repeat the experiment 10 times. First, inside the LPI matrix, which contains known association part 1 and unknown association part 0, the known association part is randomly disrupted and then divided into 5 parts, where the number of data in the last part is slightly less than that in the remaining 4 parts, and the data between every two parts are not repeated. Then 1 part is selected as the test set, and the remaining 4 parts are used as the training set, and so on, until each part of the data is used as the test set and the training set. Similarly, we take the unknown lncRNA-protein as a candidate sample, and then calculate the scores of the test sample and the candidate sample. We compare the score of each test sample with the score of the candidate sample in turn. The prediction is considered successful only when the rank of the test sample exceeds a given threshold.

The AUC values were then calculated by calculating the true positive rate TPR (sensitivity) and false positive rate FPR (specificity) for different thresholds, where sensitivity refers to the percentage of test samples above a given threshold that are positive cases and specificity refers to the percentage of pseudo-cases of lncRNA-protein associations that are below a given threshold. AUC=1 indicates that the model correctly predicted all test samples. AUC=0.5 indicates that the model is randomly predicted. AUPR refers to the area enclosed by the precision and recall versus PR curves. In these two metrics, the higher the value, the better the performance of the GILPI model, and the average value is taken as the final evaluation criterion after repeating the experiments for 10 times. The values of AUC and AUPR calculated by repeating the experiments 10 times for the GILPI model are shown in Table 2:

TABLE 2
AUC AND AUPR VALUES CORRESPONDING TO THE 5 DATA SETS

Dataset	AUC	AUPR
Data1	0.9477	0.9349
Data2	0.9496	0.9305
Data3	0.8986	0.8867
Data4	0.9706	0.8205
Data5	0.9757	0.9715
Ave.	0.9484	0.9088

As can be seen from Table 2, except for dataset 3, the AUC values of the rest of the data are all above 0.9, and dataset 5 is as high as 0.9757. Not only that, the AUPR values are also very good, except for datasets 3 and 4, which are all above 0.9, and dataset 5 is as high as 0.9715. On the whole, dataset 5 has the highest AUC and AUPR values of all datasets, and dataset 3 has the lowest AUC and AUPR values among all datasets. of all the datasets, and dataset 3 has the lowest AUC and AUPR values among all the datasets. Overall, the AUC and AUPR values of the five datasets perform very well, indicating that the GILPI model has good prediction performance.

Further, we compare the model proposed in this paper with five advanced LPI prediction models. These five models are LPI-deepGBDT [22], LPI-DLDN [23], LPI-EnANNDeep [24], LPI-EnEDT [25], and LPI-HyADBS [26], in order to measure the classification ability of the GILPI model. Their comparison results are shown in Table 3 below:

TABLE 3
COMPARISON OF AUC VALUES FOR THE 6 MODELS

Dataset	GILPI	LPI-deepGBDT	LPI-DLDN	LPI-EnANNDeep	LPI-EnEDT	LPI-HyADBS
data1	0.9477	0.9354	0.9404	0.9473	0.9297	**0.9488**
data2	0.9496	0.9423	0.9447	0.9556	0.9474	**0.9583**
data3	**0.8986**	0.8526	0.8301	0.8597	0.8235	0.8593
data4	**0.9706**	0.8542	0.9099	0.8648	0.8866	0.9162
data5	**0.9757**	0.9523	0.9302	0.9557	0.9458	0.9672
Ave.	**0.9484**	0.9074	0.9111	0.9166	0.9066	0.93

In Table 3, the GILPI model is the model proposed in this paper, and the black bolded ones are the highest values in each dataset. From the table, it can be seen that the GILPI model has the highest average AUC value of 0.9484, which was higher than LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI-HyADBS by 4.51%, 4.09%, 3.46%, 4.61%, and 1.98%, respectively. The AUCs of the GILPI model are the highest in Data 3 through Data 5, with an AUC value of 0.9757 in Data 5. In Data 1 through Data 2, the AUCs of the GILPI model are slightly lower compared to the other models at 0.9477 and 0.9496, which are only 0.11% and 0.92% lower than the highest at 0.9488 and 0.9583, but most of them are higher than the other models.

TABLE 4
COMPARISON OF AUPR VALUES FOR THE 6 MODELS

Dataset	GILPI	LPI-deepGBDT	LPI-DLDN	LPI-EnANNDeep	LPI-EnEDT	LPI-HyADBS
data1	**0.9349**	0.9043	0.9282	0.9283	0.9001	0.93
data2	0.9305	0.9242	0.9292	0.9408	0.9262	**0.9423**
data3	**0.8867**	0.8016	0.8099	0.8356	0.8005	0.8354
data4	0.8205	0.8488	0.9001	0.8683	0.8767	**0.9098**
data5	**0.9715**	0.9457	0.9246	0.954	0.9374	0.9653
Ave.	0.9088	0.8849	0.8984	0.9054	0.8882	**0.9166**

As can be seen from Table 4, the AUPR values of the GILPI model are the highest in datasets 1, 3, and 5, which are 0.9349, 0.8867, and 0.9715, respectively. with a high of 0.9715 for dataset 5, which is higher than the AUPR values of LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI- HyADBS by 2.72%, 5.07%, 1.83%, 3.64%, and 0.64%. In Data 2 and Data 4, the AUPR of the GILPI model is slightly lower 0.9305 and 0.8205. In Data 2, it is only 1.27% lower than the highest 0.9423, but all of them are higher than the LPI-deepGBDT , LPI-DLDN, and LPI-EnEDT models. In Data 5, it is only 2% lower than the highest 0.9653, but all higher than the LPI-deepGBDT, LPI-DLDN, and LPI-EnEDT models.

The five LPI prediction methods, LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI-HyADBS, are the most advanced and classical prediction models, however, the GILPI model proposed in this paper is far better than these five models. The comparative results show that the GILPI model has a powerful classification performance and is capable of mining the potential interactions between lncRNAs and proteins.

3.2 Case Studies

3.2.1 Discovery of proteins that interact with novel lncRNAs:

lncRNAs are a class of long chain RNA molecules that do not code for proteins, and he plays an important role in a variety of biological processes such as gene expression regulation and cell differentiation. In this paper, three lncRNAs, NONHSAT021830, n385685, and NONHSAT098243, which interact with 15, 16, and 19 proteins, respectively, were selected from the human dataset. In order to find out the proteins interacting with these three lncRNAs, the interaction information between the proteins associated with these three lncRNAs is masked out and these three lncRNAs are taken as the new lncRNAs in the neighboring matrix Y. Then the potential proteins are found with the GILPI modeling algorithm proposed in this paper, and the top 5 proteins predicted are shown in Table 5 below. It can be found that a total of 6 proteins were confirmed in the three datasets. Among them, in Data 1, Data 2 and Data 3, NONHSAT021830 with Q9H9S0, n385685 with Q07955, and NONHSAT098243 with P25490 were not confirmed, but they were ranked first, indicating that these three pairs of lncRNAs and proteins are likely to be associated with one another, but this is only a speculation, which needs to be further biological experiments to prove it. In summary, these results reconfirm the classification performance of GILPI. Therefore, GILPI is suitable for predicting proteins interacting with novel lncRNAs.

TABLE 5
PROTEINS INTERACTING WITH NEW LNCRNAS

Dataset	lncRNA	Protein	Confirmed	GILPI
Data1	NONHSAT021830	Q9H9S0	No	1
		P48431	No	2
		Q12968	No	3
		Q5S007	No	4
		Q8NDV7	Yes	5
Data2	n385685	Q07955	No	1
		Q9UKV8	Yes	2
		Q9UPQ9	Yes	3
		Q9HCJ0	Yes	4
		Q8NDV7	Yes	5
Data3	NONHSAT098243	P25490	No	1
		Q13285	No	2
		P60484	No	3
		Q96PU8	Yes	4
		O43251	No	5

3.2.2 Discovery of lncRNAs that interact with novel proteins:

Proteins are extremely important macromolecules in living organisms, which play key roles in physiological activities such as signaling, immune defense, cell growth and differentiation. In this paper, three proteins, O00425, Q9Y6M1 and O00425, were selected from three human datasets, which interacted with 443, 342, and 463 lncRNAs in dataset 1, dataset 2, and dataset 3, respectively. In order to find out the lncRNAs interacting with these 3 proteins, all the interaction information between the lncRNAs associated with these 3 proteins are masked out in the neighbor-joining matrix Y, and these 3 proteins are treated as new proteins to discover the potential lncRNAs with the GILPI modeling algorithm proposed in this paper. the top 5 predicted lncRNAs are shown in Table 6 below. It can be found that most of the lncRNAs were confirmed in the 3 datasets.

In dataset 1, O00425 was not confirmed with NONHSAT112460, but its ranking was 1st, indicating that O00425 and NONHSAT112460 are greatly likely to interact, but further biological proof is needed. Overall, GILPI can be used for LPI prediction of new proteins.

TABLE 6
lncRNAs THAT INTERACT WITH NOVEL PROTEINS

Dataset	Protein	lncRNA	Confirmed	GILPI
Data1	O00425	NONHSAT112460	No	1
		NONHSAT008249	No	2
		NONHSAT052575	No	3
		NONHSAT112472	No	4
		NONHSAT066972	Yes	5
Data2	Q9Y6M1	n338605	Yes	1
		n377669	Yes	2
		n345648	Yes	3
		n381041	Yes	4
		n342241	Yes	5
Data3	O00425	NONHSAT016408	Yes	1
		NONHSAT093392	No	2
		NONHSAT124481	Yes	3
		NONHSAT041141	Yes	4
		NONHSAT025390	Yes	5

3.2.3 Finding new LPIs based on known LPIs:

Immediately after that, based on the known LPIs, we use the model GILPI proposed in this paper to discover new LPIs. the top 50 lncRNA-protein pairs with the highest scores on the five datasets are filtered as shown below, where the circle represents the lncRNA, the hexagon represents the protein, the ones with known associations are connected by a solid line, the ones with unknown associations are connected by a dashed line, and the ones connected by a light blue color with a light green color are with known associations, and yellow and light green are connected with unknown associations, and these top 50 contain lncRNA-protein pairs with known associations and unknown associations.

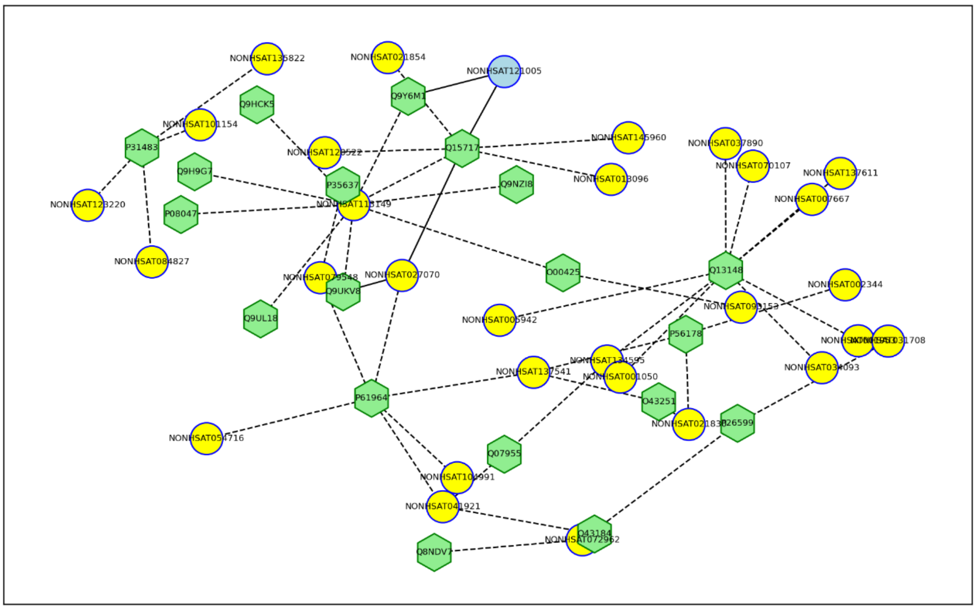


FIGURE 4: Top 50 lncRNA-protein pairs with the highest scores in Data 1

In Data 1, there are a total of 55,165 lncRNA-protein pairs. In the calculated top 50, there are a total of 4 lncRNA-protein pairs with known associations and 46 pairs with unknown associations, e.g., NONHSAT113149 is associated with Q15717 and NONHSAT137541 is associated with P61964 with unknown associations, but these two lncRNA-protein pairs are ranked as the 2nd and 3rd out of 55165 pairs, so there is a high probability that they are interacting, but this needs to be proved by further biological experiments.

In Data 2, there are a total of 74,340 lncRNA-protein pairs, and there are 44 pairs of unknown associations in the top 50 calculated as n385685 with O14746, n385685 with O95793, etc., but they are all ranked very high, and even the 2 pairs of unknown associations, n385685 with O14746 and n385685 with O95793, ranked 1st and 3rd, respectively, inside the 74340 pairs ranked 1st and 3rd respectively, so there is a great possibility that these 2 pairs of lncRNAs and proteins are interacting. In the 6 known association pairs, their rankings are 2, 14, 18, 29, 45, 46 respectively.

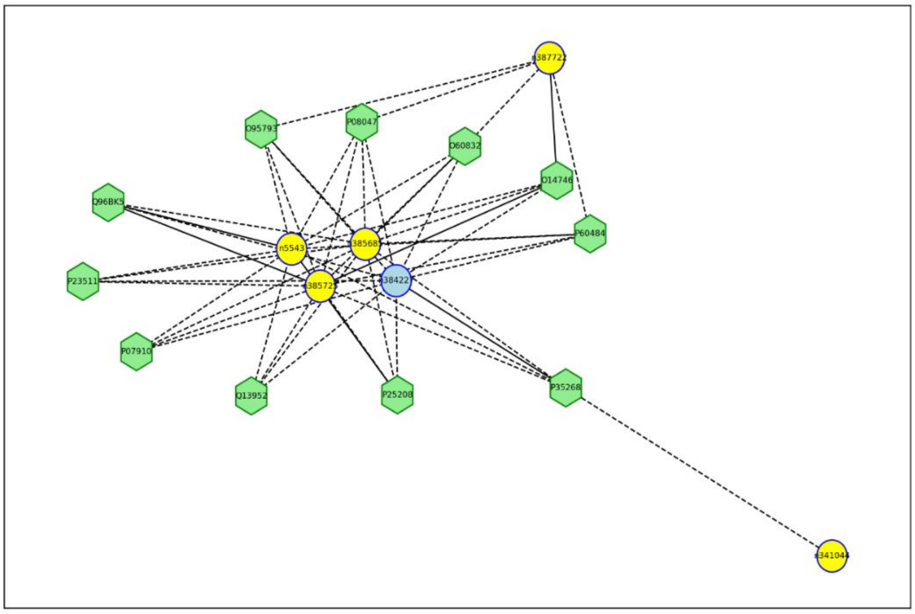


FIGURE 5: Top 50 lncRNA-protein pairs with the highest scores in Data 2

In Data 3, there are a total of 26,730 pairs of lncRNAs and proteins, with 4 pairs of known associations and 46 pairs of unknown associations in the top 50 calculated. Although the number of unknown associations is high, they are all ranked highly. For example, NONHSAT121765 with P60484 and NONHSAT121765 with O43251 are ranked 1st and 4th, respectively, for these unknown associations, and there may be interactions between them, but this is only a guess, and further experiments are needed to prove it.

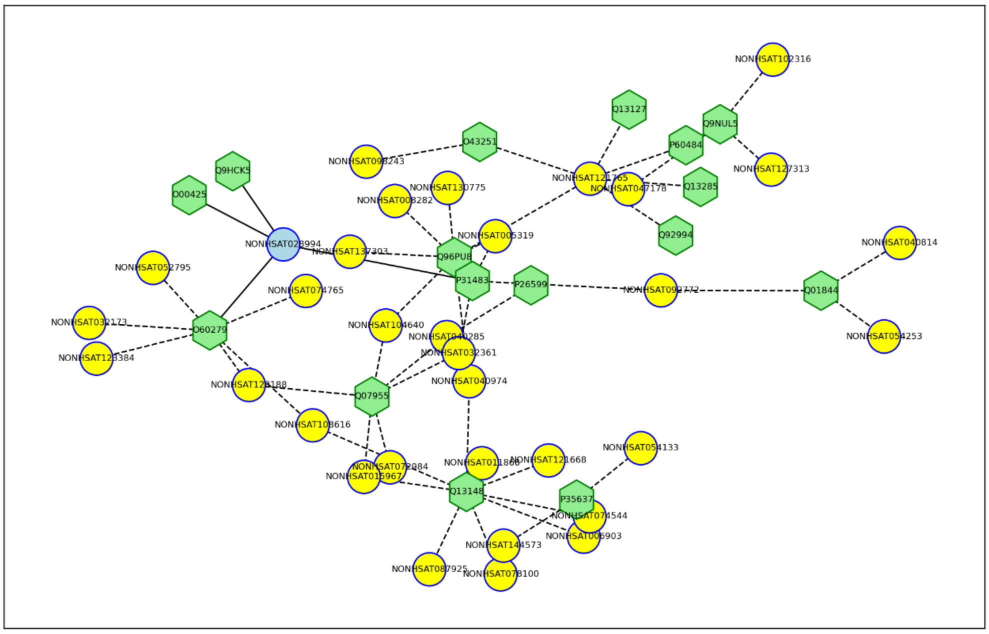


FIGURE 6: Top 50 lncRNA-protein pairs with the highest scores in Data 3

In Data 4, there are a total of 3,815 lncRNA-protein pairs, and among the top 50 calculated pairs, there are 30 pairs with known associations and 20 pairs with unknown associations, for example, the 3 unknown associations of AthlncRNA229 with 15229884, AthlncRNA227 with 79326195, and AthlncRNA302 with 18423684 , ranked 8th, 11th and 15th inside all 3815 lncRNA-protein pairs, suggesting that there may be interactions between these lncRNA-protein pairs.

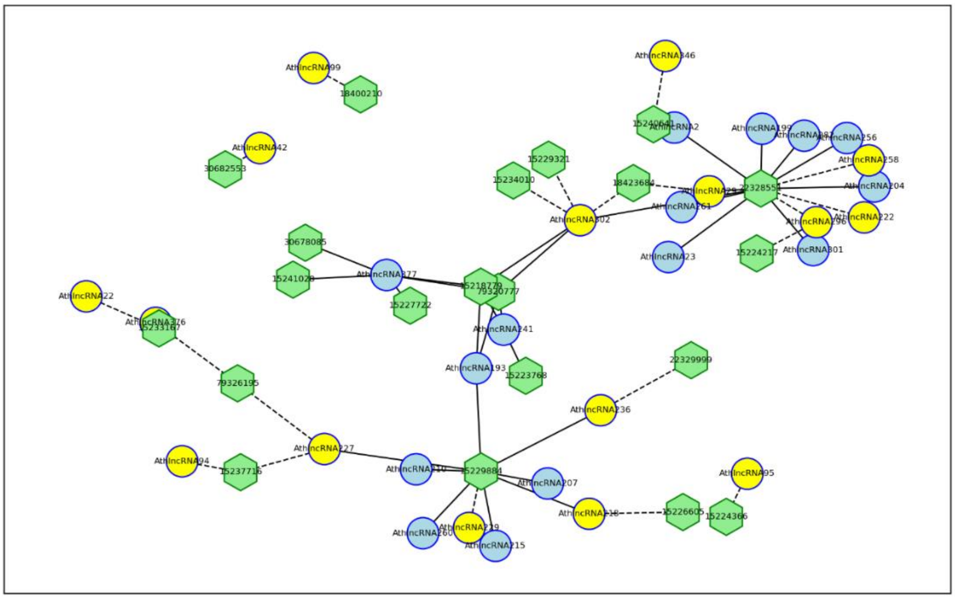


FIGURE 7: Top 50 lncRNA-protein pairs with the highest scores in Data 4

In dataset 5, there are a total of 22,133 pairs of lncRNAs with proteins, which is the most in the five datasets. In the top 50 calculated pairs, there are 24 pairs with known associations and 26 pairs with unknown associations. The known associations are basically ranked at the top of the 50 pairs, and the unknown associations are ranked a little bit later. However, the ranking of unknown associations is also very good in all the 22,133 pairs. For example, ZmalncRNA1062 with B4FLX0 and ZmalncRNA1263 with C0PF88 ranked 9th and 15th respectively, so these two pairs and other unknown associations of lncRNAs and proteins in the 50 pairs, they may interact with each other.

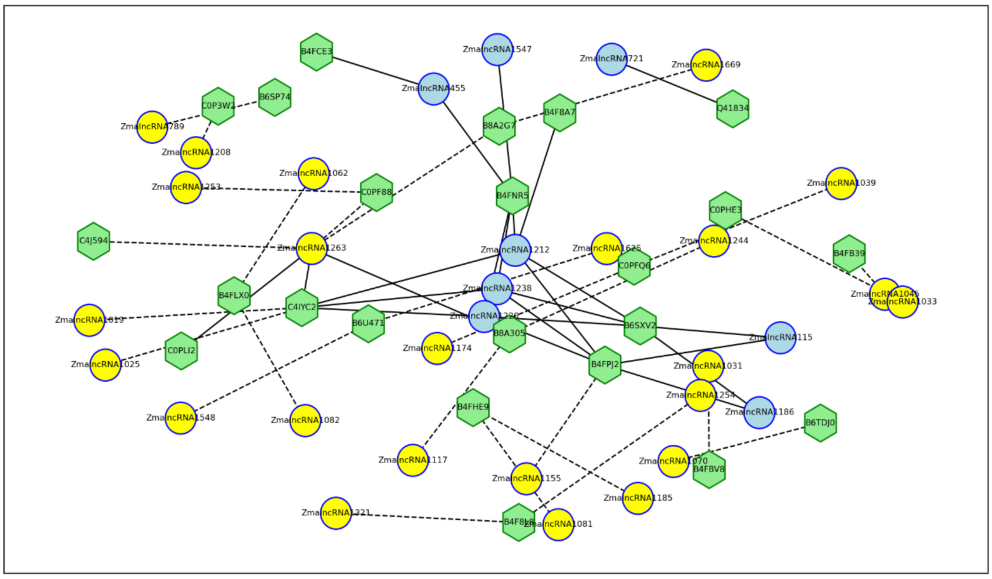


FIGURE 8: Top 50 lncRNA-protein pairs with the highest scores in Data 5

IV. DISCUSSION

Characterization of lncRNA-protein interaction relationships helps to discover the function and mechanism of action of lncRNAs. In this paper, we developed a prediction model GILPI incorporating Pearson similarity, Jaccard similarity to classify lncRNA-protein interaction relationships. The experiment was repeated 10 times to train the model using five-fold cross-validation and compared with other state-of-the-art LPI prediction models. The experimental results show that the GILPI prediction model proposed in this paper is able to classify lncRNA-protein interaction relationships more accurately and can be used to discover new LPIs.

Under the five-fold cross-validation, most of the performances of the five prediction models, LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI-HyADBS, are much lower than that of the GILPI model proposed in this paper. For training, among the five data, after randomly disrupting the known associations, 80% is selected as the training set, 20% as the test set, and the remaining unknown associations as the candidate samples, and then the test set and the candidate samples are scored and ranked. In addition, it is further shown in the case study that the GILPI prediction model proposed in this paper can mine useful information for new lncRNAs or new proteins.

The GILPI prediction model proposed in this paper demonstrates a powerful LPI classification capability. It incorporates Pearson similarity and Jaccard similarity to fully mine the complex biological information between lncRNA-protein, and then utilizes the characteristics of Graphlet interaction direct connection and indirect connection on lncRNA-protein network to deeply mine the hidden features between lncRNA and protein. It greatly enriches the features when the model is trained, and makes the prediction performance of the model more accurate and powerful. Although the GILPI model can accurately identify new LPIs, it also has the following problems: one is that the network-based method has a defect that it cannot predict separate lncRNAs and proteins, so the GILPI model proposed in this paper can not predict single lncRNAs and proteins. Second, the Graphlet interactions used in this paper have the number of nodes within 4 nodes, so the information beyond 4 nodes is ignored, resulting in insufficiently rich training features obtained. Third, the time complexity of this model is high. It takes a long time for the model to run once, and repeating the experiment 10 times in this paper takes a lot of time.

V. CONCLUSIONS

lncRNAs play a crucial role in many biological activities, such as gene transcription, translation and other processes. Not only that, lncRNAs also affect numerous diseases, so recognizing the lncRNA and protein interaction relationship can be a good grasp of the biological function of lncRNAs, which is important for the treatment of disease therapy, diagnosis and so on.

First, five datasets were collected; second, features of lncRNAs and proteins were extracted from the sequence data using pyfeat and BioTriangle, respectively. Third, these features were analyzed by Pearson's correlation coefficient to calculate the similarity between lncRNAs and the similarity between proteins. Fourth, the Jaccard similarity between lncRNAs and proteins was calculated based on the LPI network, and then the corresponding Pearson similarity and Jaccard similarity were averaged to construct the lncRNA-lncRNA similarity network and protein-protein similarity network. The experiment was repeated 10 times, and GILPI was compared with five state-of-the-art LPI prediction methods, namely, LPI-deepGBDT, LPI-DLDN, LPI-EnANNDeep, LPI-EnEDT, and LPI-HyADBS, and the results showed that the GILPI prediction model had a strong LPI classification performance.The GILPI prediction model in the case study also achieved good results.

In future studies, we will first integrate various lncRNA and protein related datasets from different data sources. Secondly, mining the secondary and tertiary structures of proteins fused into lncRNA-protein pairs makes it possible to predict the relationship between a single lncRNA-protein pair. Then secondly, other nodes than the four nodes are considered in Graphlet interactions to make the acquired features more complete and rich. Finally, the computational efficiency is optimized by utilizing high-performance computing resources such as GPU acceleration and distributed computing to reduce the time of a single run, developing more efficient algorithms to handle large-scale datasets with less computational redundancy, and optimizing and automating the tuning of the model parameters to reduce the time needed to manually adjust the parameters.

ACKNOWLEDGEMENTS

This work was supported by Science and Technology Planning Project of Guizhou Province of China (No. Qian Ke He Ji Chu -ZK[2021] Yi Ban 315), Science Foundation of Guizhou University of Finance and Economics (No. 2021KYYB21).

REFERENCES

[1] KHALIL A M, RINN J L. RNA–protein interactions in human health and disease [J]. Seminars in Cell & Developmental Biology, 2011, 22(4): 359-65.

[2] Tiwari A, Srivastava R. A survey of computational intelligence techniques in protein function prediction. Int J Proteomics. 2014;2014: 845479.

[3] Tamang S, Acharya V, Roy D, Sharma R, Aryaa A, Sharma U, Khandelwal A, Prakash H, Vasquez KM, Jain A (2019) Snhg12: an lncRNA as a potential therapeutic target and biomarker for human cancer. Front Oncol 9:901.

https://doi.org/10.3389/fonc.2019.00901

[4] Mao Z, Li H, Du B, Cui K, Xing Y, Zhao X, Zai S (2017) LncRNA dancr promotes migration and invasion through suppression of lncRNA-let in gastric cancer cells. Biosci Rep. https://doi.org/10.1042/BSR20171070

[5] Batista PJ, Chang HY. Long noncoding RNAs: cellular address codes in development and disease. Cell. 2013;152(6):1298–307.

[6] Selth LA, Gilbert C, Svejstrup JQ. RNA immunoprecipitation to determine RNA–protein associations in vivo. Cold Spring Harbor Potoc. 2009;2009(6):pdb–prot5234.

[7] Liu H, Ren G, Hu H, Zhang L, Ai H, Zhang W, Zhao Q (2017) Lpi-nrlmf: lncrna-protein interaction prediction by neighborhood regularized logistic matrix factorization. Oncotarget. https://doi.org/10.18632/oncotarget.21934

[8] Zhang T, Wang M, Xi J, Li A (2018) Lpgnmf: predicting long non-coding RNA and protein interaction using graph regularized nonnegative matrix factorization. IEEE/ACM Trans Comput Biol Bioinform 17(1):189–197.

https://doi.org/10.1109/TCBB.2018.2861009

[9] Ma Y, He T, Jiang X (2019) Projection-based neighborhood nonnegative matrix factorization for lncRNA-protein interaction prediction. Front Genet 10:1148. https://doi.org/10.3389/fgene.2019.01148

[10] Zhao Q, Yu H, Ming Z, Hu H, Ren G, Liu H (2018) The bipartite network projection-recommended algorithm for predicting long non-coding RNA-protein interactions. Mol Therapy Nucleic Acids 13:464–471. https://doi.org/10.1016/j.omtn.2018.09.020

[11] Ge M, Li A, Wang M (2016) A bipartite network-based method for prediction of long non-coding RNA-protein interactions. Genom Proteom Bioinform 14(1):62–71. https://doi.org/10.1016/j.gpb.2016.01.004

[12] Jia L, Luan Y. Multi-feature Fusion Method Based on Linear Neighborhood Propagation Predict Plant LncRNA-Protein Interactions. Interdiscip Sci. 2022 Jun;14(2):545-554. doi: 10.1007/s12539-022-00501-7. Epub 2022 Jan 17. PMID: 35040094.

[13] Li A, Ge M, Zhang Y, Peng C, Wang M (2015) Predicting long noncoding RNA and protein interactions using heterogeneous network model. BioMed Res Int. https://doi.org/10.1155/2015/ 671950

[14] Yuan J, Wu W, Xie C, Zhao G, Zhao Y, Chen R (2014) Npinter v2. 0: an updated database of ncrna interactions. Nucleic Acids Res 42(D1):D104–D108. https://doi.org/10.1093/nar/gkt1057

[15] Xie C, Yuan J, Li H, Li M, Zhao G, Bu D, Zhu W, Wu W, Chen R, Zhao Y (2014) Noncodev4: exploring the world of long noncoding RNA genes. Nucleic Acids Res 42(D1):D98–D103. https:// doi.org/10.1093/nar/gkt1222

[16] Consortium U (2019) Uniprot: a worldwide hub of protein knowledge. Nucleic Acids Res 47(D1):D506–D515. https://doi.org/10. 1093/nar/gky1049

[17] Zheng X, Wang Y, Tian K, Zhou J, Guan J, Luo L, Zhou S (2017) Fusing multiple protein-protein similarity networks to efectively predict lncRNA-protein interactions. BMC Bioinform 18(12):11– 18. https://doi.org/10.1186/s12859-017-1819-1

[18] Zhang W, Qu Q, Zhang Y, Wang W (2018) The linear neighborhood propagation method for predicting long non-coding RNAprotein interactions. Neurocomputing 273:526–534. https://doi. org/10.1016/j.neucom.2017.07.065

[19] Bai Y, Dai X, Ye T, Zhang P, Yan X, Gong X, Liang S, Chen M (2019) PLNCRNADB: a repository of plant LNCRNAS and LNCRNA-RBP protein interactions.Curr Bioinform 14(7):621–627. https://doi.org/10.2174/1574893614666190131161002

[20] R. Muhammod, S. Ahmed, D. Md Farid, S. Shatabda, A. Sharma, and A.Dehzangi, “PyFeat: A python-based effective feature generation tool for DNA, RNA and protein sequences,” Bioinformatics, vol. 35, no. 19, pp. 3831–3833, 2019.

[21] J. Dong et al., “Biotriangle: A web-accessible platform for generating various molecular representations for chemicals, proteins, DNAs/RNAs and their interactions,” J. Cheminformat., vol. 8,no. 1, pp. 1–13, 2016.

[22] Zhou L, Wang Z, Tian X, Peng L. LPI-deepGBDT: a multiple-layer deep framework based on gradient boosting decision trees for lncRNA-protein interaction identification. BMC Bioinformatics. 2021 Oct 4;22(1):479. doi: 10.1186/s12859-021-04399-8. PMID: 34607567; PMCID: PMC8489074.

[23] Peng L, Wang C, Tian X, Zhou L, Li K. Finding lncRNA-Protein Interactions Based on Deep Learning With Dual-Net Neural Architecture. IEEE/ACM Trans Comput Biol Bioinform. 2022 Nov-Dec;19(6):3456-3468. doi: 10.1109/TCBB.2021.3116232. Epub 2022 Dec 8. PMID: 34587091.

[24] Peng L, Tan J, Tian X, Zhou L. EnANNDeep: An Ensemble-based lncRNA-protein Interaction Prediction Framework with Adaptive k-Nearest Neighbor Classifier and Deep Models. Interdiscip Sci. 2022 Mar;14(1):209-232. doi: 10.1007/s12539-021-00483-y. Epub 2022 Jan 10. PMID: 35006529.

[25] Peng L, Yuan R, Shen L, Gao P, Zhou L. LPI-EnEDT: an ensemble framework with extra tree and decision tree classifiers for imbalanced lncRNA-protein interaction data classification. BioData Min. 2021 Dec 3;14(1):50. doi: 10.1186/s13040-021-00277-4. PMID: 34861891; PMCID: PMC8642957.

[26] Zhou L, Duan Q, Tian X, Xu H, Tang J, Peng L. LPI-HyADBS: a hybrid framework for lncRNA-protein interaction prediction integrating feature selection and classification. BMC Bioinformatics. 2021 Nov 26;22(1):568. doi: 10.1186/s12859-021-04485-x. PMID: 34836494; PMCID: PMC8620196.

Machinability Investigations on Al6063+TiO_2 Metal Matrix Material

K Udayani[1*], S Gajanana[2], P Laxminarayana[3], B Ravikumar[4]

[*1]Research Scholar, Department of Mechanical Engineering, Osmania University, Hyderabad, India
[2]Professor, Department of Mechanical Engineering, MVSREC, Hyderabad, India
[3]Professor, Department of Mechanical Engineering, Osmania University, Hyderabad, India
[4]Assistant Professor, Department of Mechanical Engineering, MVSREC, Hyderabad, India
*Corresponding Author

Received: 03 July 2024/ Revised: 10 July 2024/ Accepted: 16 July 2024/ Published: 31-07-2024

Abstract— *Investigations into the machinability of Al6063 alloy reinforced with TiO_2 particles typically focus on understanding how the addition of TiO_2 affects the machining characteristics of the metal matrix material compared to the base Al6063 alloy. Here are some aspects that such studies would typically explore: Tool Wear, cutting forces, Material removal rate, surface roughness. These investigations are crucial for understanding how the addition of TiO_2 particles modifies the machinability of Al6063 alloy and for optimizing machining processes to ensure efficient production of components with desirable mechanical and surface properties.*

Keywords— DoE, MRR, Process parameters, Resultant Force.

I. Introduction

Aluminium alloy 6063, often referred to simply as Al6063, is a popular metal matrix alloy known for its excellent combination of mechanical properties and workability. Al6063 primarily consists of aluminium (Al) as the base metal, with significant additions of magnesium (Mg) and silicon (Si). Typical composition ranges are approximately: Al: 97.0 - 98.5%, Mg: 0.45 - 0.9%, Si: 0.2 - 0.6%

Other trace elements like iron (Fe), titanium (Ti), Zinc (Zn) and chromium (Cr) in small amounts.

II. Literature

Abdul Nazeer et al.'s study [1] attempted to examine the impact of alumina Al_2O_3 when reinforced with aluminium 6063 matrix. The composite was prepared using the liquid metallurgical approach (Stir Casting Technique), with the reinforcement varying from 0 to 8wt% in increments of 2wt%. Research on prepared composite systems includes mechanical, wear, fractography, and X-ray diffraction, as well as testing carried out in accordance with ASTM and ISO standards. Following the discovery that the reinforcement was distributed uniformly throughout the matrix alloy, the mechanical test revealed that the mechanical properties, such as hardness, toughness, and tensile strength, improved with an increase in contain reinforcement. A similar finding was made in the wear test, where an increase in contain reinforcement led to improved wear resistance. A fractured tensile specimen was examined under a scanning electron microscope. Reactions in Al–10 weight percent TiC metal matrix composites have been studied by A.R. Kennedy et al. [2]. Samples were heated between 600 and 900°C for 48 hours and then held at 700°C for up to 240 hours. The composition, shape, and amounts of the reaction phases present have been determined using X-ray diffraction, scanning electron microscopy, and image analysis. By varying the percentage of the reinforced element alumina in the base matrix alloy Al 6063, Bhavana Mathur et al. [3] focused on the mechanical properties of the metal matrix composite, such as tensile behavior, hardness, and surface characteristics of Al 6063/Al2O3 Alumina reinforced metal matrix composites. The samples prepared by stir casting process by varying the percentage of alumina in the base matrix alloy Al 6063 were tested for finding the ultimate tensile strength, followed by hardness and surface characteristics.

Ersan Aslan, et al., [4] conducted an experimental study to achieve this by employing Taguchi techniques. Combined effects of three cutting parameters, namely cutting speed, feed rate and depth of cut on two performance measures, flank wear (VB) and surface roughness (Ra), were investigated employing an orthogonal array and the analysis of variance (ANOVA). Optimal cutting parameters for each performance measure were obtained; also the relationship between the parameters and the performance measures were determined using multiple linear regression. To reduce surface roughness (Ra and Rz), İlhan Asiltürk, et al. [5] concentrated on optimizing turning parameters based on the Taguchi method. A CNC turning machine's L9 orthogonal array has been used in experiments. Tests for dry turning are performed using coated carbide cutting tools on hardened AISI 4140 (51 HRC). K. Hemalatha et al. [6] studied the stir casting technique, which is used to cast Al 6063 plates with different masses of Al_2O_3 (3%, 6%, and 9%). In addition, the material's mechanical properties, such as tensile strength and hardness, are tested, and the distribution of aluminium and alumina is investigated through microstructure analysis and hardness distribution. The impact of alumina volume percentage and solution heat-treatment on the corrosion behavior of Al (6063) composites and its monolithic alloy in basic and acidic environments was examined by K. K. Alaneme et al. [7]. Using two-step stir casting, Al (6063) - Al_2O_3 particulate composites with volumes of 6, 9, 15, and 18 percent alumina were created. Mass loss and corrosion rate measurements were utilized as criteria for evaluating the corrosion behavior of the composites. According to M. Amrutha Pavani et al. [8], a meager effort would be made to create silicon carbide particulate MMCs based on aluminium with the goal of creating a traditional, low-cost technique of generating MMCs and achieving uniform dispersion of ceramic material. In order to accomplish these goals, a two-step stir casting procedure has been suggested, and a property study has since been conducted. SiC particles and aluminium 6063 T6 have been selected as the matrix and reinforcing materials, respectively. The weight fraction of SiC will be varied in experiments (in 5% steps) while all other parameters will remain constant. Tests for Hardness and Impact (including microstructure) would be used to evaluate the outcomes for this "development method." Al-MMC has been created by combining 5wt% ZrO2 and Al2O3 reinforcement into the Al6063 aluminium alloy matrix, according to Munmun Bhaumik et al.'s [9] research. The process of stir casting has been used to create MMC. X-ray diffraction analysis and scanning electron microscopy (SEM) have been used to characterize the prepared casted MMC. For the manufactured MMC, measurements have been made of its mechanical (hardness, tensile test, bend test, and compression test) and physical (density) characteristics. Analysis of the fracture surface has been done. Al-MMC fractures are found to be brittle in nature. Tests for Hardness and Impact (including microstructure) would be used to evaluate the outcomes for this "development method." The creation of multi-phase hybrid composites made of polyester reinforced with E-glass fiber and ceramic particles was reported by S.S. Mahapatra et al. [10]. It also looks at how these composites respond to erosion and wear. Finally, it compares the effects of three distinct particle fillers—silicon carbide (SiC), alumina (Al_2O_3), and cement by-pass dust (CBPD)—on the wear properties of glass-polyester composites. To do this, Taguchi's orthogonal arrays are used in the design of experiments approach to create the erosion test schedule for an air jet type test rig. The Taguchi technique makes it possible to identify the ideal parameter combinations that minimize the rate of erosion. W.H. Yang, et al., [11] employed the Taguchi approach, a strong tool to design optimization for quality, is used to identify the ideal cutting parameters for turning operations. An orthogonal array, the signal-to-noise (S/N) ratio, and the analysis of variance (ANOVA) are applied to evaluate the cutting properties of S45C steel bars employing tungsten carbide cutting tools.

III. EXPERIMENTATION

3.1 Introduction:

The following composition of Al6063 was used based on strength criteria, and the same material is used for this experimentation by reinforcing TiO_2 in varying percentages 2% and 6% prepared by using stir casting method. The dimensions of the work piece after machining are, length is 180mm and diameter is 22mm. Conducted trials on a lathe with a casted workpiece and HSS single point cutting tool. Optimum composition of Al 6063 alloy having highest tensile strength is shown in tables 1 below:

TABLE 1
Weight Percentage of metals in Al6063

Metal	Al	Mg	Si	Fe	Cu	Zn	Ti	Mn	Cr
Weight %	98.65	0.45	0.2	0.3	0.1	0.1	0.05	0.05	0.1

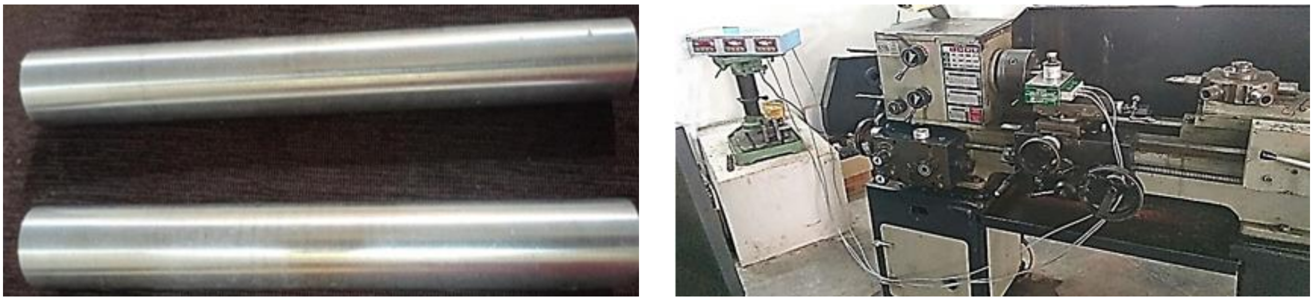


FIGURE 1: Work piece and experimental setup

Taguchi L8 Orthogonal Array is used for conduct of machining trails. Table 2 represents the number of trials considering two levels for each factor.

TABLE 2
Taguchi Design Matrix

Trail No	s (rpm)	f (mm/rev)	d (mm)	r (^{0})
1	150	0.21	0.2	15
2	150	0.21	0.5	20
3	150	0.421	0.2	20
4	150	0.421	0.5	15
5	445	0.21	0.2	20
6	445	0.21	0.5	15
7	445	0.421	0.2	15
8	445	0.421	0.5	20

Where **s** is cutting speed(rpm), **f** is feed(mm), **d** is depth of cut(mm) and **r** is rake angle of tool (in degrees)

3.2 Material Removal Rate (MRR):

It is a key metric in manufacturing and machining processes that indicates the volume of material removed from a workpiece over a given period.

The formula to calculate Material Removal Rate is:

$$\mathrm{MRR} = \frac{w_1 - w_2}{\mathrm{t}}$$

Where, w_1 is the weight of the workpiece before machining (gm) w_2 is the weight of the workpiece after machining (gm) tm is the machining time (min).

From the experimental investigation, the following tabular calculation for MRR has been developed.

TABLE 3
MRR respected to Al6063+TiO_2 (2%)

Trail No.	s (rpm)	f (mm/rev)	d (mm)	r (^{0})	w1	w2	t (min)	MRR (gm/min) $\frac{w_1 - w_2}{t}$
1	150	0.21	0.2	15	135	134.142	0.54	1.588
2	150	0.21	0.5	20	134.142	131.788	0.511	4.606
3	150	0.421	0.2	20	131.788	130.977	0.233	3.48
4	150	0.421	0.5	15	130.977	130.069	0.233	3.896
5	445	0.21	0.2	20	130.069	127.799	0.171	13.27
6	445	0.21	0.5	15	127.799	126.736	0.167	6.365
7	445	0.421	0.2	15	126.736	126.222	0.076	6.763
8	445	0.421	0.5	20	126.222	125.628	0.103	5.766

TABLE 4
MRR respected to Al6063+TiO_2 (6%)

Trail No.	s (rpm)	f (mm/rev)	d (mm)	r (^{0})	w1	w2	t (min)	MRR (gm/min) $\frac{w_1 - w_2}{t}$
1	150	0.21	0.2	15	136	134.402	0.467	3.421
2	150	0.21	0.5	20	134.402	133.377	0.495	2.07
3	150	0.421	0.2	20	133.377	132.792	0.231	2.532
4	150	0.421	0.5	15	132.792	131.522	0.246	5.162
5	445	0.21	0.2	20	131.522	130.587	0.15	6.233
6	445	0.21	0.5	15	130.587	129.538	0.152	6.901
7	445	0.421	0.2	15	129.538	128.89	0.072	9
8	445	0.421	0.5	20	128.89	127.662	0.082	14.975

Dynamometer is used in to measure machining forces. Based on the recorded values, the resulting forces are computed and are listed in tables 5 and 6 for each case, respectively.

TABLE 5
Force response of Al6063+TiO_2(2%)

Trail No.	t (min)	F_x (Kgf)	F_y (Kgf)	Resultant Force RF (kgf)
1	0.54	0	5	5
2	0.511	13	30	32.695
3	0.233	0	7	7
4	0.233	3	11	11.401
5	0.171	10	23	25.079
6	0.167	3	9	9.486
7	0.076	1	2	2.224
8	0.103	0	3	3

TABLE 6
Force response of Al6063+TiO_2(6%)

Trail No.	t (min)	F_x (Kgf)	F_y (Kgf)	Resultant Force RF (kgf)
1	0.467	7	18	19.313
2	0.495	0	6	6
3	0.231	0	4	4
4	0.246	5	20	20.61
5	0.15	0	7	7
6	0.152	3	8	8.544
7	0.072	0	6	6
8	0.082	4	4	14.566

IV. DEVELOPMENT OF A MATHEMATICAL MODEL

A statistical technique called Taguchi design of experiments is used to create effective trials that optimize procedures and end products with the least amount of experimentation possible.

4.1 Taguchi Design for Al6063+TiO_2(2%):

Taguchi Orthogonal Array Design

L8(2^4), Factors: 4, Runs: 8

Regression Equation for MRR:

MRR = -2.90 + 0.01576 s - 7.02 f - 3.72 d + 0.425 r

Regression Equation for RF:

RF = -5.5 - 0.0138 s - 57.6 f + 14.4 d + 1.98 r

4.2 Taguchi Design for Al6063+TiO_2(6%):

Taguchi Orthogonal Array Design

L8(2^4), Factors: 4, Runs: 8

Regression Equation for MRR:

MRR = -8.09 + 0.02027 s + 15.45 f + 6.60 d + 0.066 r

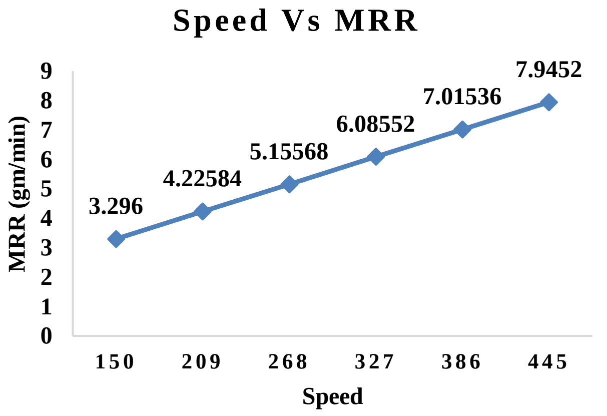
Regression Equation for RF:

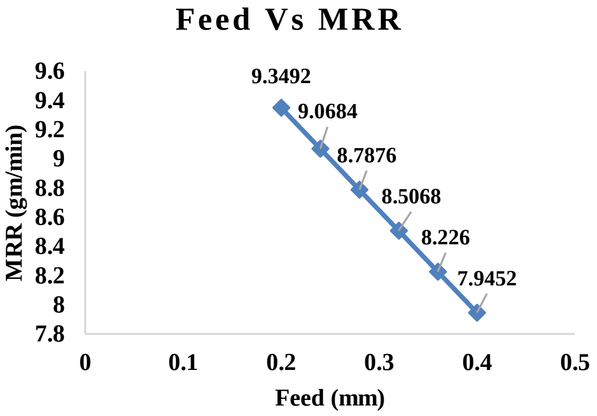
RF = 28.8 - 0.0117 s + 5.1 f + 11.2 d - 1.15 r

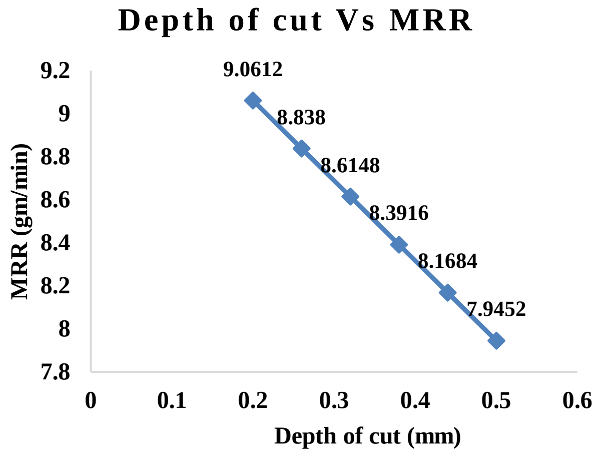
V. GRAPHICAL ANALYSIS

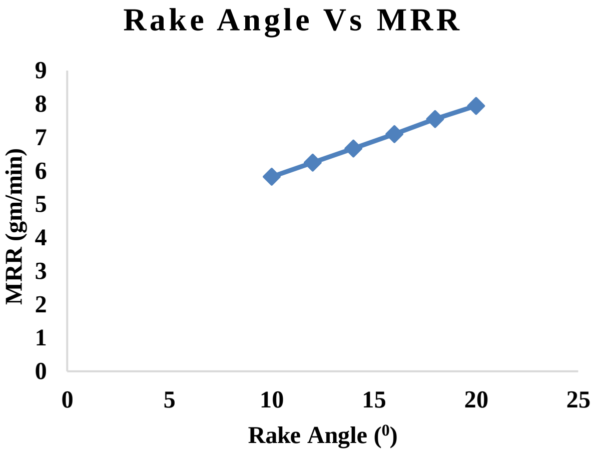
Following graphs shows the relationship between each parameter, the material removal rate, and the resultant force for Al6063+ TiO_2 (2%) and Al6063+ TiO_2 (6%).

5.1 Graphs Obtained for Al6063+TiO_2(2%):

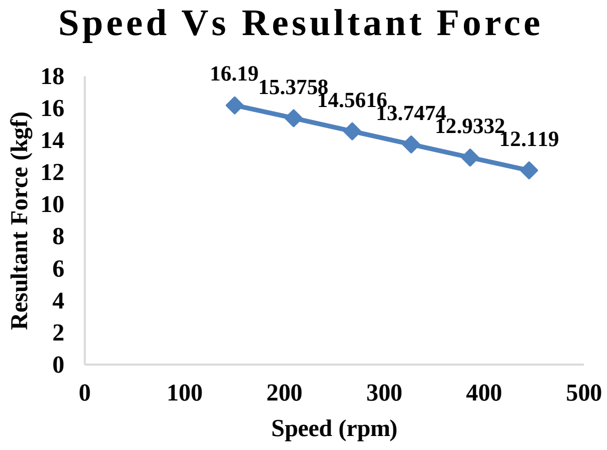


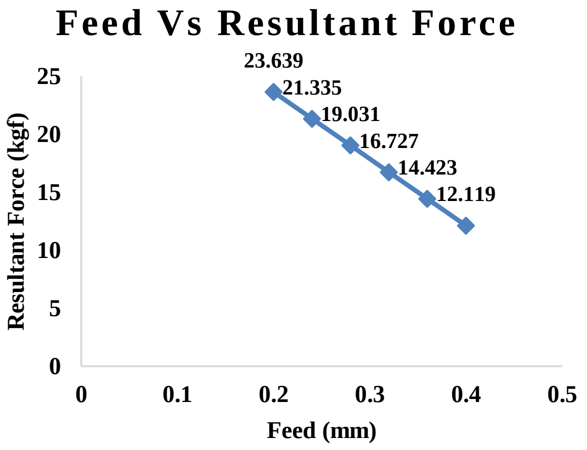


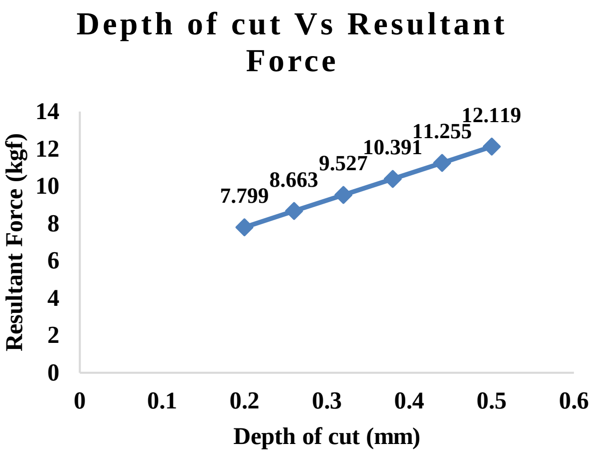


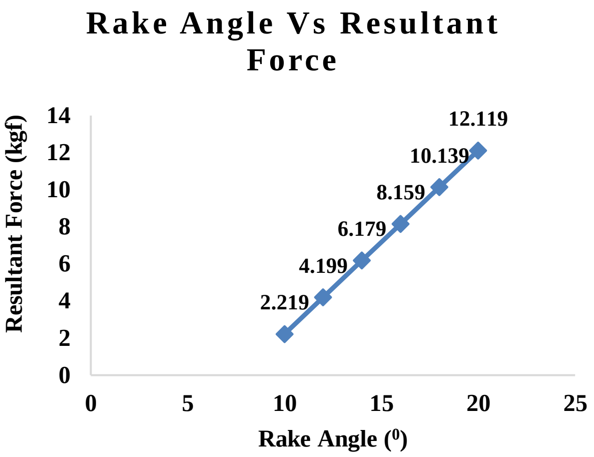


5.2 Graphs Obtained for Al6063+TiO_2(2%) for RF:

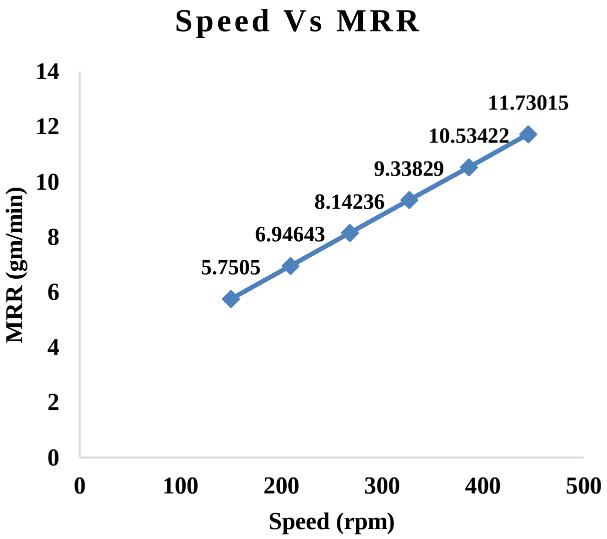


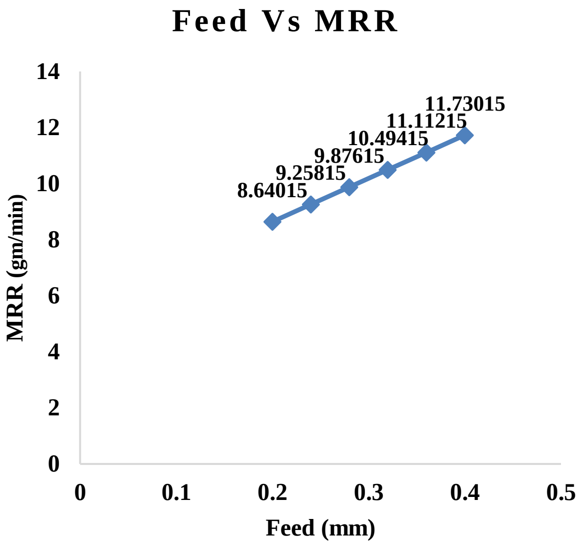


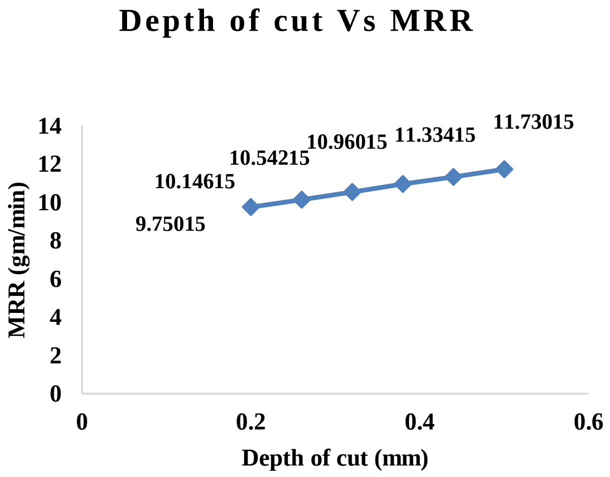

Depth of cut Vs Resultant Force

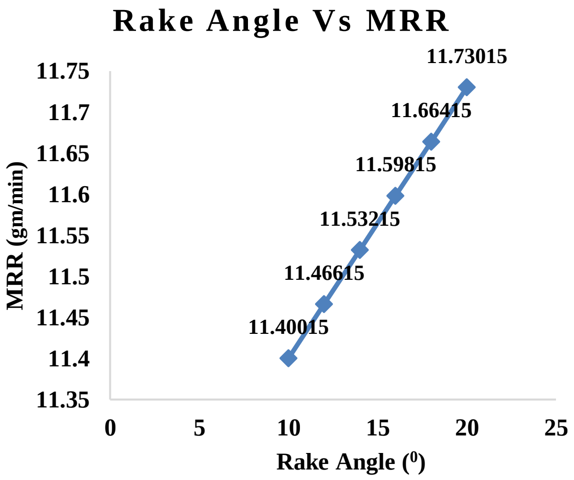

Rake Angle Vs Resultant Force

5.3 Graphs Obtained for Al6063+TiO_2(6%) for MRR:

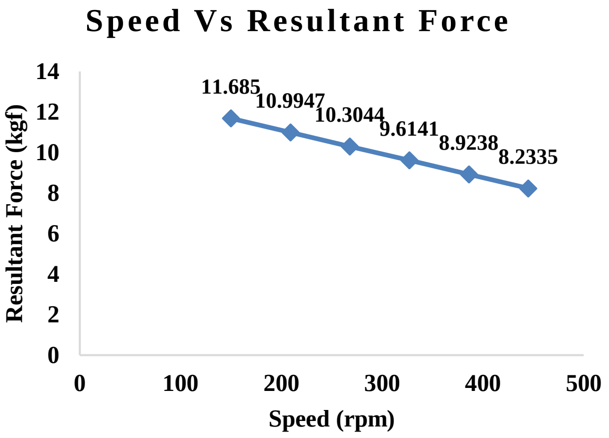

Speed Vs MRR

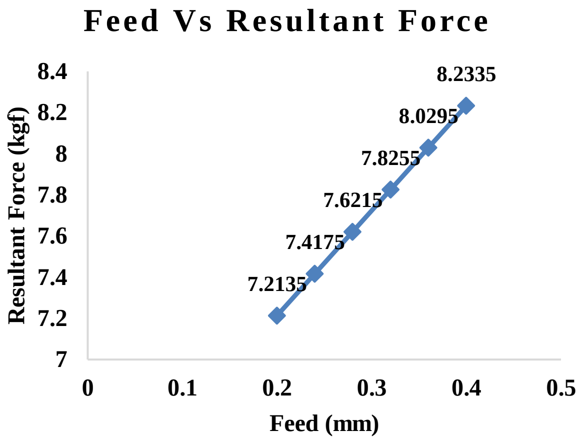

Feed Vs MRR

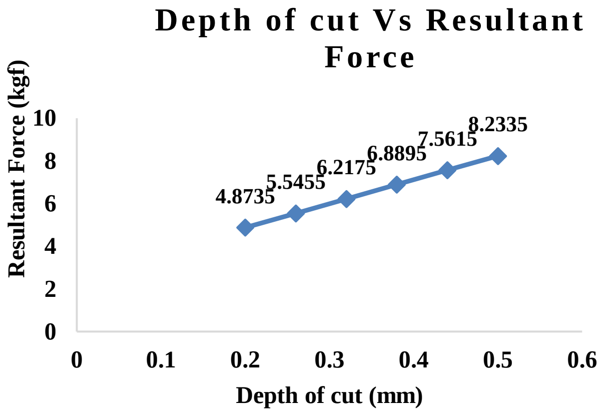

Depth of cut Vs MRR

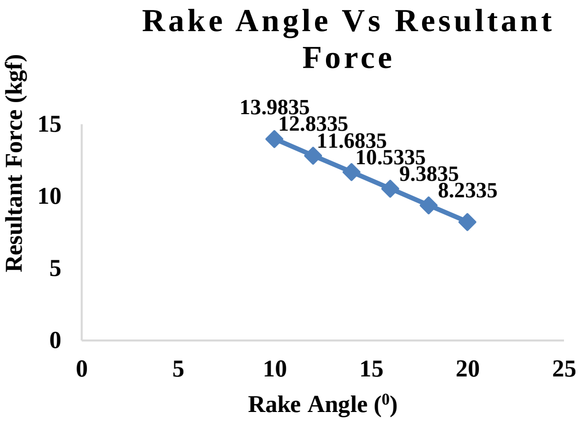

Rake Angle Vs MRR

5.4 Graphs Obtained for Al6063+TiO$_2$(6%) for RF:


Speed Vs Resultant Force


Feed Vs Resultant Force


Depth of cut Vs Resultant Force


Rake Angle Vs Resultant Force

VI. CONCLUSIONS

The following conclusions are drawn from the work carried out

- The maximum material removal rate observed is 13.27 gm/min for trail-5 in case of TiO$_2$(2%), for the machining parameters s = 445 rpm, f = 0.21 mm, d =0.2mm, at r= 200 and corresponding cutting force is 25.079 kgf.
- The maximum material removal rate observed is 14.975 gm/min for trial-8 in case of TiO$_2$(6%), where the machining parameters are s = 445 rpm, f = 0.421 mm, mm, d =0.5mm, at r= 20° and corresponding cutting force is 14.566 kgf.
- From the graphs it is evident that in case of machining Al6063 with TiO$_2$ (2%) increase in speed and rake leads to higher MRR whereas for depth of cut and feed shows decline trend.
- Resultant in case of machining Al6063 with TiO$_2$ (2%) is increasing for increase in depth of cut and rake angle, however it has down trend in case of increase in speed and feed.
- Al6063 with TiO$_2$ (6%) machining has uptrend behaviour of MRR with all input parameters
- While machining Al6063 with TiO$_2$ (6%) resultant for has uptrend with depth of cut and feed but has downtrend with speed and rake angle.

REFERENCES

[1] Abdul Nazeer, Mir Safiulla, “Mechanical and Wear Properties of Al6063 Metal Matrix Composite Reinforced with Al2O3 Particles”, International Journal of Recent Technology and Engineering (IJRTE), ISSN: 2277-3878, Volume-8 Issue-5, January 2020

[2] A.R. Kennedy, D.P. Weston, M.I. Jones, “Reaction in Al–TiC metal matrix composites”, Elsevier Ltd., Materials Science and Engineering: A Volume 316, Issues 1–2, 15 November 2001, Pages 32-38

[3] Bhavana Mathur, Prashant Kumar, “Mechanical Characterization of Stir Casted Al 6063/Al2O3 (Alumina) Reinforced Metal Matrix Composites”, International Journal of Technical Research & Science (Special Issue) ISSN No.:2454-2024 (online) ICRDET-2019

[4] Ersan Aslan a, Necip Camuşcu a, Burak Birgören, “Design optimization of cutting parameters when turning hardened AISI 4140 steel (63 HRC) with Al2O3 + TiCN mixed ceramic tool”, Materials & Design, Volume 28, Issue 5, 2007, Pages 1618-1622

[5] İlhan Asiltürk, Harun Akkuş, “Determining the effect of cutting parameters on surface roughness in hard turning using the Taguchi method”, Measurement, Volume 44, Issue 9, November 2011, Pages 1697-1704

[6] K.Hemalatha, V. S. K.Venkatachalapathy, N.Alagumurthy, “Processing and Synthesis of Metal Matrix Al 6063/Al2o3 Metal Matrix Composite by Stir Casting Process”, Int. Journal of Engineering Research and Applications, ISSN : 2248-9622, Vol. 3, Issue 6, Nov-Dec 2013, pp.1390-1394 Nov-Dec 2013

[7] K. K. Alaneme and M. O. Bodunrin, “Corrosion Behavior of Alumina Reinforced Aluminium (6063) Metal Matrix Composites”, Journal of Minerals & Materials Characterization & Engineering, Vol. 10, No.12, pp.1153-1165, 2011

[8] M.Amrutha Pavani, M.Venkaiah, “Mechanical Properties of Stir Casted Al6063- Sic Metal Matrix Composite”, Proceedings of International Conference on Recent Trends in Mechanical Engineering-2K15

[9] Munmun Bhaumik and Kalipada Maity, “Fabrication and characterization of the Al6063/5%ZrO2 /5% Al2O3 composite”, National Conference on Processing and Characterization of Materials IOP Publishing IOP Conf. Series: Materials Science and Engineering 178 (2017)

[10] S.S. Mahapatra a, Amar Patnaik, “Study on mechanical and erosion wear behavior of hybrid composites using Taguchi experimental design”, Elsevier Ltd., Materials & Design Volume 30, Issue 8, September 2009, Pages 2791-2801

[11] W.H. Yang, Y.S. Tarng, “Design optimization of cutting parameters for turning operations based on the Taguchi method”, Journal of Materials Processing Technology, Volume 84, Issues 1–3, 1 December 1998, Pages 122-129.

Hybrid Intelligence: DT-CNN's Solution to Credit Card Fraud Detection

Anjalika Arora[1*], Jinguo Lian[2]

[*1]Managerial Economics and Computer Science Student, University of Massachusetts Amherst, USA
[2]Department of Mathematics, University of Massachusetts Amherst, USA
*Corresponding Author

Received: 04 July 2024/ Revised: 11 July 2024/ Accepted: 20 July 2024/ Published: 31-07-2024

Abstract— *The proliferation of electronic transactions has heightened the vulnerability to credit card fraud, demanding more robust detection methodologies. This paper introduces DT-CNN, an innovative hybrid model that integrates a Decision Tree (DT) and a Convolutional Neural Network (CNN) to enhance the accuracy* and efficiency of fraud detection significantly. By leveraging decision trees' interpretability and CNNs' pattern recognition capabilities, DT-CNN offers a comprehensive approach to identifying fraudulent transactions. Unlike conventional models, DT-CNN adeptly addresses challenges related to precision* and recall*, achieving notable performance metrics in real-world datasets prone to biases. The hybrid model's architecture enables effective learning from vast and intricate datasets. This study builds upon previous research by advancing techniques in feature engineering, dataset balancing, and overfitting mitigation, positioning DT-CNN as a dependable solution for combating fraud. Detailed insights into its architecture, training methodology, and performance evaluation further underscore DT-CNN's effectiveness in combating credit card fraud.*

Keywords— Convolutional Neural Network, Credit Card Fraud Detection, Decision Trees.

I. INTRODUCTION

The ubiquity of electronic transactions in modern society has brought unprecedented convenience but has also given rise to a significant surge in credit card fraud, imposing substantial financial burdens on consumers and financial institutions. According to recent studies, the global cost of credit card fraud exceeded $32 billion in 2021 alone, with projections indicating a further upward trend [1]. Traditional fraud detection methods, often reliant on static rule-based systems, have proven inadequate in addressing the evolving tactics employed by fraudsters, necessitating innovative and adaptive solutions [2].

In response to this pressing challenge, this paper introduces an innovative hybrid model that combines a decision tree with a convolutional neural network to enhance credit card fraud detection capabilities. Our proposed model leverages the strengths of both traditional machine learning and advanced deep learning techniques, aiming to improve the accuracy and efficiency of fraud detection. Additionally, inspired by principles of credit risk analysis, such as the probability of default, our model offers a comprehensive approach to identifying fraudulent transactions, thereby reducing potential financial losses.

Recent statistics underscore the urgency of developing powerful fraud detection mechanisms. Newly released Federal Trade Commission data shows that consumers reported losing more than $5.8 billion to fraud in 2021, an increase of more than 70 per cent over the previous year [4]**.** Additionally, the average cost of a fraudulent transaction rose to approximately $3 for every $1 of fraud, further highlighting the financial ramifications of inadequate fraud prevention measures [3].

While standalone approaches, such as CNN or decision tree, have demonstrated efficacy in certain contexts, they often exhibit limitations when deployed independently. CNN, renowned for its prowess in pattern recognition, may lack interpretability, hindering its adoption in sensitive financial domains. Conversely, the decision tree offers transparency in decision-making but may struggle to capture intricate patterns inherent in transactional data. Furthermore, despite achieving high accuracy, both models independently can suffer from poor precision and recall, leading to a high number of false positives and false negatives. This is particularly problematic in real-life scenarios where the average cost of misclassifications is extremely high, as highlighted by the aforementioned financial ramifications.

By amalgamating these methodologies into a hybrid framework, our model seeks to reconcile these shortcomings, providing financial institutions with a comprehensive and adaptable solution for combating credit card fraud. Through rigorous experimentation and statistical analysis, we demonstrate the superiority of our hybrid model over standalone approaches, showcasing its enhanced accuracy, precision, recall, and overall performance. By harnessing the complementary strengths of decision trees and CNNs, enhanced by feature engineering techniques, our model represents a significant advancement in credit card fraud detection, offering effectiveness and reliability in safeguarding against fraudulent activities.

The subsequent sections of this paper are structured as follows: Section 2 reviews related work in the field of credit card fraud detection, providing context and background for our approach. Section 3 describes the design methodology, detailing the individual components of the Decision Tree and CNN models, and their integration into the DT-CNN hybrid model. It also elaborates on the DT-CNN hybrid model, including the dataset used, preprocessing steps, and the combined training process. Section 4 presents the results of our experiments, comparing the performance of the Decision Tree, CNN, and DT-CNN hybrid models. Section 5 discusses the implications of our findings, analyzing the strengths and limitations of each model. Finally, Section 6 concludes the paper, by summarizing our contributions and highlighting the significance of the DT-CNN hybrid model in enhancing credit card fraud detection. At the end of the paper, an appendix is included to provide definitions for technical terms used in this paper.

II. RELATED WORK

The major works related to credit card fraud detection using Machine Learning are presented below.

TABLE 1
RELATED WORK [5]

No.	Title of the Research Paper	Features Extraction	Model	Weaknesses
1	Credit Card Fraud Detection in Payment Using Machine Learning Classifiers [6]	Taking advantage of the properties of algorithms to extract important features	Naïve Bayes, C4.5 Decision Tree, Bagging Ensemble Learning	The research paper fails to address feature engineering and imbalance in the dataset
2	A machine learning based credit card fraud detection using the GA algorithm for feature selection [7]	Synthetic Minority Oversampling Technique (SMOTE) method	Decision Tree (DT), Random Forest (RF), Logistic Regression (LR), Artificial Neural Network (ANN), and Naive Bayes (NB)	Despite using one of the dataset normalization methods, the results suffer from overfitting
3	Credit Card Fraud Detection Using Artificial Neural Network [8]	None	Artificial Neural Network (ANN), support vector machines (SVM), k-nearest neighbors (KNN)	Does not use dataset balancing techniques, which might lead to unreliable results
4	Digital payment fraud detection methods in digital ages and Industry 4.0 [9]	Undersampling and feature reduction method using principal components analysis	Logistic regression (LR), decision tree (DT), k-nearest neighbors (KNN), random forest (RF), and autoencoder	The effects of undersampling and oversampling vary across algorithms, impacting prediction accuracy and reliability.
5	Detection of Fraudulent Transactions in Credit Card using Machine Learning Algorithms [10]	None	Artificial Neural Networks, Decision Trees, Support Vector Machine, Logistic Regression, and Random Forest	There are weaknesses in the architectures of the algorithms used: ANN performance is influenced by the hardware architecture. Decision Tree (DT) suffers from overfitting. Support Vector Machines (SVM) require longer training times for larger datasets. Random Forests (RF) are excessively sensitive to data with diverse values and attributes.
6	Auto Loan Fraud Detection using Dominance-based Rough Set Approach versus Machine Learning Methods [11]	The ADASYN method is employed to achieve a balanced dataset	Logistic regression, random forest, k-nearest neighbors, naive Bayes, multilayer perceptron, AdaBoost, quadrant discriminative analysis, pipelining and ensemble learning	Accuracy varied for the categories of the dataset used, as the models recorded low accuracy for fraud transactions, indicating that the method used to balance the dataset is not appropriate
7	Credit Card Fraud Detection Using CNN [12]	Convolutional Neural Networks (CNNs), SMOTE	CNN	High computational cost, potential for overfitting on highly imbalanced datasets and only used accuracy as the evaluation metric

III. DESIGN METHODOLOGY

3.1 Decision Tree and CNN- A brief:

3.1.1 Decision Tree Classifier:

Decision Tree Classifiers are widely used machine learning algorithms for classification tasks. They work by recursively splitting the data into subsets based on the value of input features, forming a tree structure where each internal node represents a test on a feature, each branch represents the outcome of the test, and each leaf node represents a class label. Decision Tree is known for its simplicity, interpretability, and ability to handle both numerical and categorical data.

The features (X) and target variable (y) are separated, and the data is split into training and testing sets using a 70-30 ratio to ensure a sufficient portion for model validation. The Decision Tree Classifier from scikit-learn is used, initialized with the 'entropy' criterion* to measure the quality of splits and with a fixed random state (any number can be used) to ensure reproducibility. The classifier is trained on the training data, and predictions are made on the test set. The model's performance is evaluated using key metrics, including accuracy, precision, recall, and F1 score, to provide a comprehensive understanding of its effectiveness. Additionally, a confusion matrix* is generated to detail the true positives, true negatives, false positives, and false negatives, and this matrix is visualized using a heatmap for clearer insights into the model's performance. This methodology ensures a thorough and effective approach to training and evaluating the Decision Tree model for fraud detection.

3.1.2 Convolutional Neural Networks:

Convolutional Neural Networks are a powerful class of deep learning models predominantly used for tasks involving image analysis, but they are also applicable to sequential data such as time series. Binary cross entropy, also known as log loss, is a loss function used in binary classification tasks to measure the difference between probability distributions, particularly between predicted probabilities and actual binary outcomes (0 or 1).

CNN operates by leveraging convolutional layers to automatically learn spatial hierarchies of features from input data. Each convolutional layer applies learnable filters across the input data, extracting local patterns. Subsequent layers, such as pooling layers, reduce dimensionality while retaining important features. Fully connected layers at the end of the network combine high-level features for classification.

Features (X) and the target variable (y) were separated, followed by a split into training and testing sets using a 70-30 ratio to ensure robust model validation. The neural network architecture was constructed using TensorFlow and Keras, utilizing three Conv1D* layers for feature extraction, followed by two MaxPooling1D* layers for dimensionality reduction and then finally two Dense* layers for classification, where the final dense layer acts as the output layer. The model was compiled with 'rmsprop' optimizer* and 'binary cross entropy'* loss function, optimized for binary classification of fraud detection. Training occurred over 5 epochs, with validation against the test set to assess performance metrics such as accuracy, precision, recall, and F1 score. Post-training, predictions were made on the test data, and evaluation metrics were computed using scikit-learn functions. A confusion matrix was generated to detail true positives, true negatives, false positives, and false negatives, visualized using Seaborn. This structured approach ensures a comprehensive evaluation of the CNN model's efficacy in fraud detection tasks.

3.2 DT-CNN Hybrid Model:

Decision Tree provides explicit rules for classification, making it easy to understand how decisions are made. Following this, a Convolutional Neural Network model is constructed due to its ability to automatically learn and extract relevant features from raw data through its hierarchical and layered structure, reducing the need for manual feature engineering.

To improve the overall model performance, the strengths of both models are combined. Hence a new DT-CNN model is proposed. This hybrid approach aims to enhance the metrics by leveraging the interpretability of Decision Tree and the feature learning capabilities of CNN, leading to a quicker and more accurate fraud detection system.

The dataset was initially split into features (X) and the target variable (y). The data was then divided into training and testing sets using a 70-30 ratio, ensuring sufficient data for model validation while preserving the integrity of class distribution. A Decision Tree Classifier from scikit-learn was instantiated with the 'entropy' criterion to assess the split quality and a fixed random state for reproducibility. The classifier was trained on the training data using the fit method, acquiring the ability to classify transactions based on input features. Predictions were subsequently generated for the test set using the predict method.

Following this, misclassified predictions were identified by comparing the predicted labels from the Decision Tree model with the actual labels in the test set. The instances corresponding to misclassified predictions were extracted from the test data. This subset of misclassified data was then standardized using Standard Scaler for compatibility with the Convolutional Neural Network model.

The CNN model was similarly constructed to the previous CNN model using TensorFlow and Keras, comprising Conv1D layers for feature extraction and Dense layers for classification. It was compiled with the 'rmsprop' optimizer and 'binary_crossentropy' loss function, tailored for binary classification tasks in fraud detection. The model was trained on the training data for 5 epochs, and its performance was evaluated using the misclassified data extracted from the Decision Tree predictions. Predictions from the CNN model on this subset of misclassified data were computed and evaluated using standard metrics to assess its efficacy in correctly identifying previously misclassified fraudulent transactions. This method is also summarized in the flowchart in Table 1.

This methodology integrates the strengths of both Decision Tree and CNN models, leveraging the Decision Tree's initial predictions to refine and test the CNN's performance specifically on instances where the initial classifier faltered. This iterative approach aims to enhance overall fraud detection accuracy by focusing CNN's learning on challenging cases identified by the Decision Tree classifier.

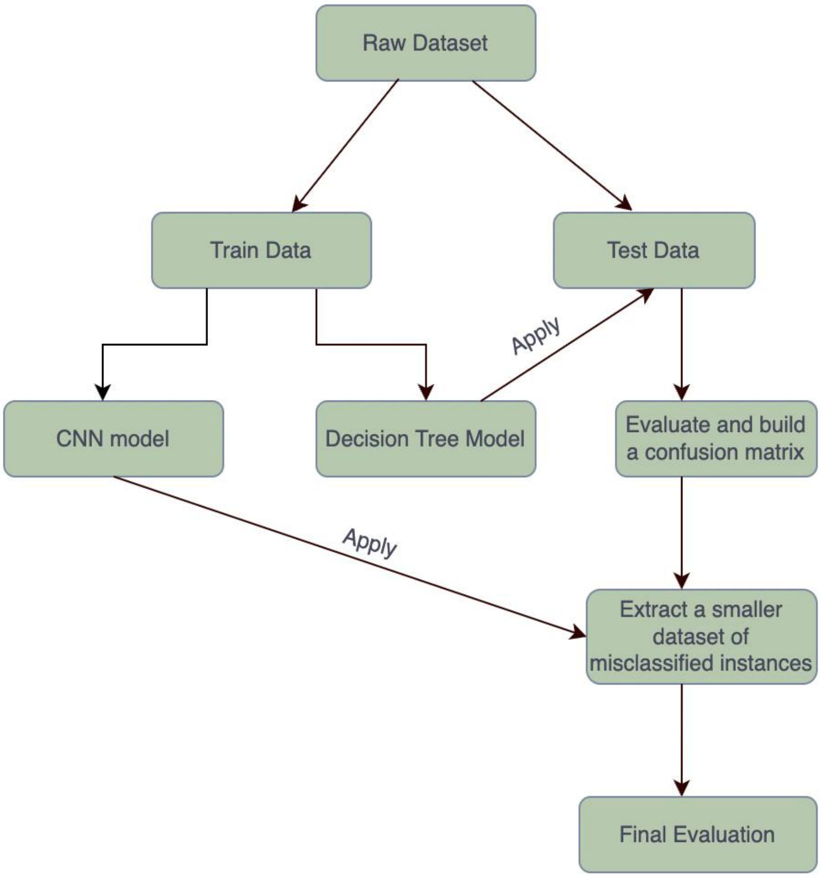

Raw Dataset
Train Data
Test Data
Apply
CNN model
Decision Tree Model
Evaluate and build a confusion matrix
Apply
Extract a smaller dataset of misclassified instances
Final Evaluation

FIGURE 1. Flow chart of DT-CNN process

3.3 Dataset and Data Pre-Processing:

The study utilizes a Kaggle credit card fraud detection dataset comprising 284,807 transactions, with 492 being fraudulent. The dataset has 31 features and 2 labels, where 0 is the label for non-fraudulent transactions and 1 is the label for fraudulent transactions. Moreover, the feature names have been removed to maintain the security of sensitive data stored in the CSV.

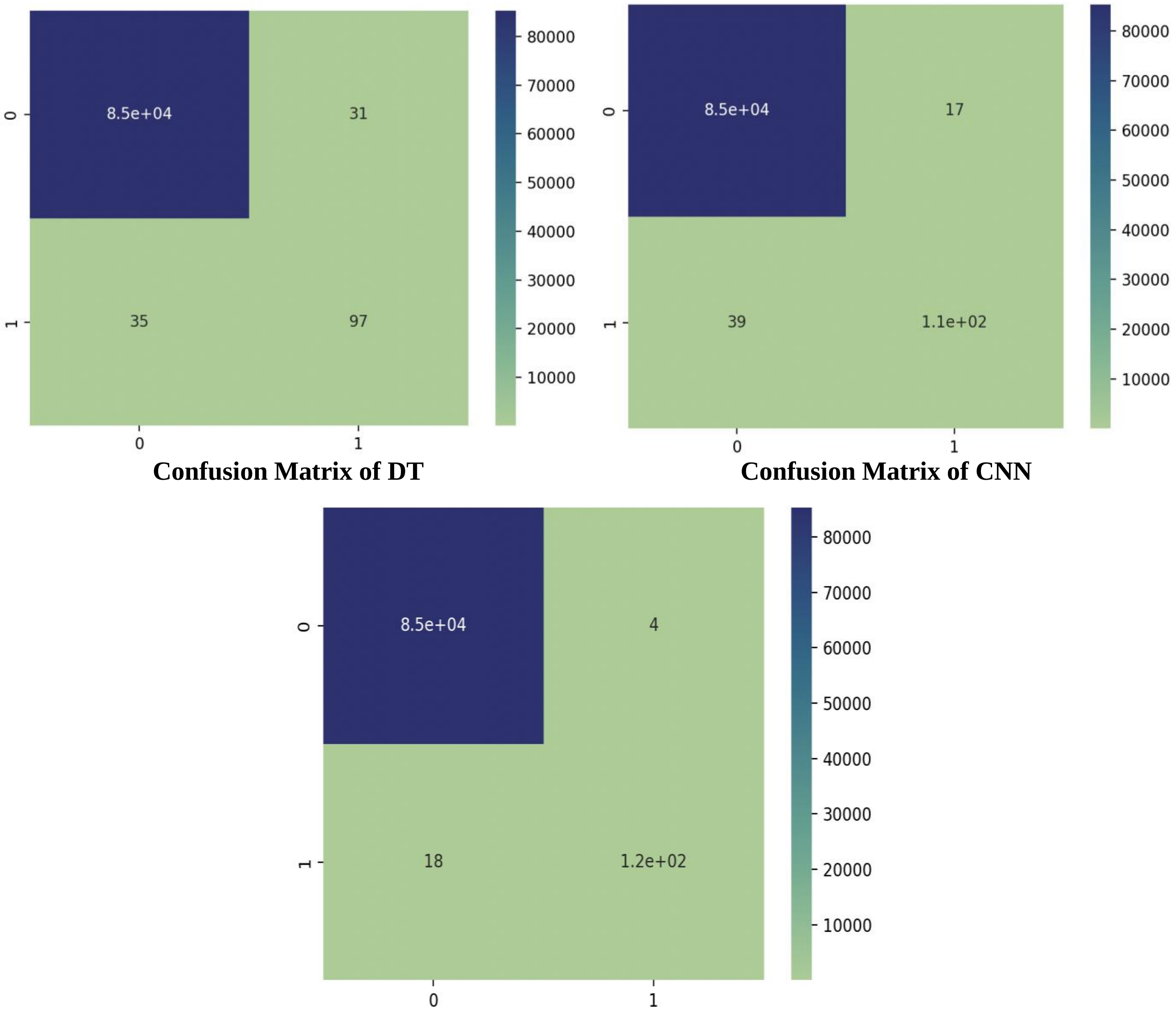
To handle this imbalanced data, we preprocess the dataset by cleaning and preparing it using Pandas and NumPy libraries. Decided not to use SMOTE as the results were still suffering from overfitting [8]. The dataset is divided into training and testing sets using a 70-30 split, ensuring sufficient data for model training and validation. Lastly, scikit-learn's Standard Scaler is employed to normalize the features, ensuring they are on a similar scale between -1 and 1.

IV. RESULTS

The results of testing all 3 of the models are summarized in Table 2 and Figure 2

TABLE 2
TABLE SUMMARIZING THE EVALUATION METRICS OF ALL MODELS

Model/Metric	Accuracy	Precision	Recall	F-1 Score
DT	0.9992	0.7578	0.7349	0.7462
CNN	0.9994	0.8651	0.7365	0.7956
DT-CNN	0.9997	0.9995	0.9999	0.9997



Confusion Matrix of DT

Confusion Matrix of CNN

Confusion Matrix of DT-CNN

FIGURE 2: Comparison of Confusion Matrices

(Top-left: Actual non-fraudulent Predicted non-fraudulent, Top-right: Actual non-fraudulent Predicted fraudulent, Bottom-left: Actual fraudulent Predicted non-fraudulent, Bottom-right: Actual fraudulent Predicted fraudulent)

V. DISCUSSION

The accuracy metric shows that all three models achieve very high accuracy, with the DT-CNN Hybrid model performing the best.

Reasons for High Accuracy:

- **Imbalanced Dataset:** The dataset is heavily skewed towards non-fraudulent transactions. Since the majority class dominates, even a simple model can achieve high accuracy by correctly predicting the majority class most of the time.
- **Decision Tree Classifier:** This model achieves high accuracy by correctly classifying most non-fraudulent transactions. However, its performance is limited by its difficulty in detecting the minority class.
- **CNN Model:** The CNN model, with its ability to capture complex patterns, slightly improves accuracy by better identifying fraudulent transactions.
- **DT-CNN Hybrid Model:** This model further enhances accuracy by combining the strengths of both the Decision Tree and CNN. The initial decision tree classification followed by CNN refinement on misclassified instances ensures that even difficult cases are handled effectively.

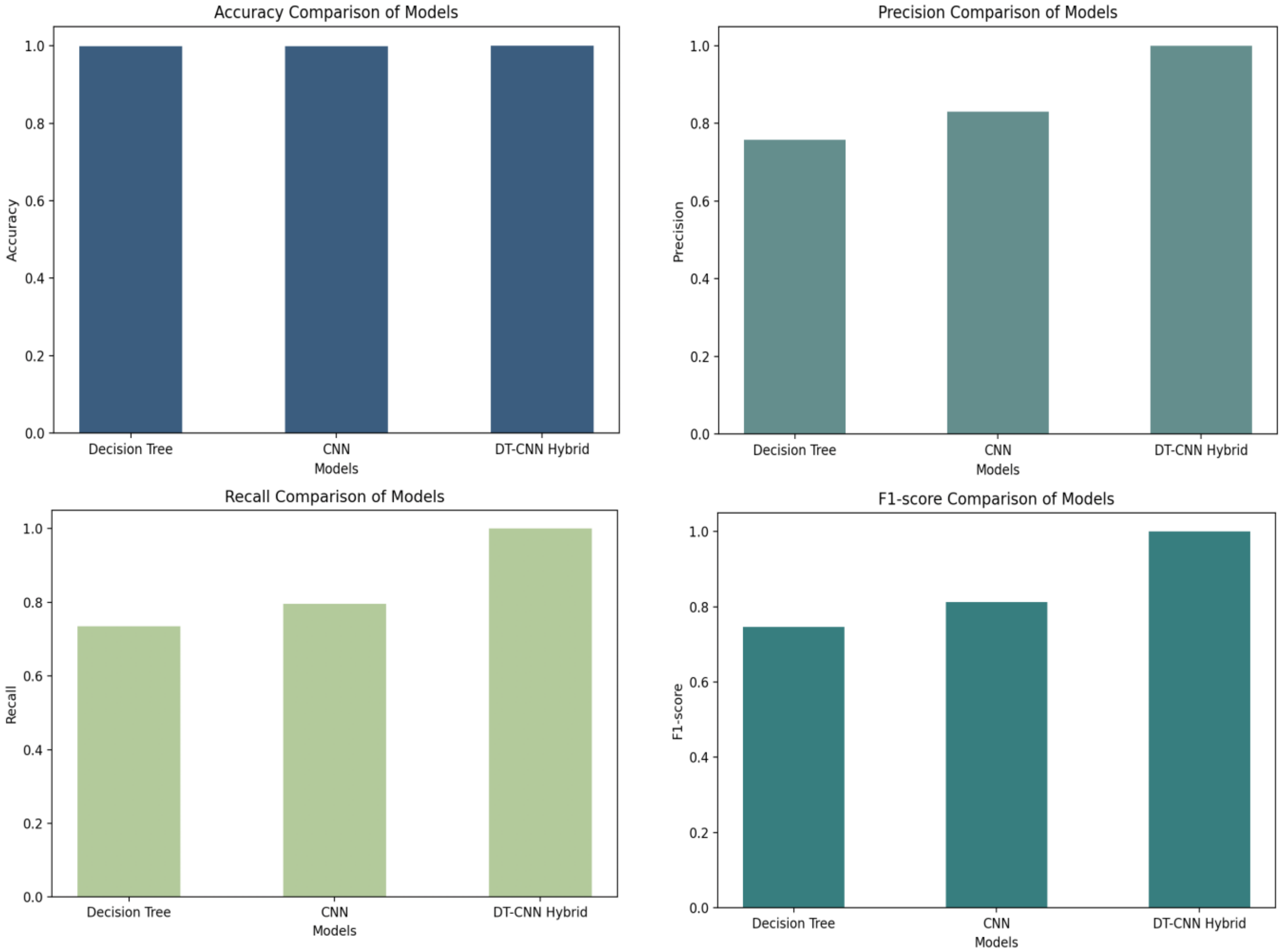


FIGURE 3: Comparison of the evaluation metrics for the 3 model

The results of the study underscore the strengths and limitations of the three models—Decision Tree (DT) Classifier, Convolutional Neural Network (CNN), and the DT-CNN Hybrid model—in the context of credit card fraud detection.

5.1 Decision Tree Classifier:

The Decision Tree Classifier achieves high accuracy with low precision, recall and F-1 score*, primarily driven by its ability to correctly identify the majority class (non-fraudulent transactions). Its simplicity and interpretability make it a favorable choice, especially for understanding the decision-making process. However, the inherent imbalance in the dataset, where non-fraudulent transactions vastly outnumber fraudulent ones, skews the performance metrics. This imbalance results in lower precision and recall for detecting fraudulent transactions, indicating room for improvement. The model's susceptibility to overfitting, especially with deep trees, further highlights the need for techniques such as pruning or ensemble methods (e.g., Random Forests or Gradient Boosting) to enhance robustness and generalization.

5.2 Convolutional Neural Network:

The CNN model demonstrates a balanced performance, with high accuracy and mediocre precision, recall, and F1-score, indicating its proficiency in identifying fraudulent transactions while maintaining a low false-positive rate. Its ability to automatically learn and extract complex patterns from raw data is a significant advantage. However, like the Decision Tree Classifier, CNN's high accuracy is influenced by the imbalanced dataset. The model excels in classifying the majority class but still faces challenges in improving precision and recall for the minority class (fraudulent transactions). Techniques such as data augmentation, oversampling the minority class, or utilizing a more balanced dataset could further enhance the model's performance.

5.3 DT-CNN Hybrid Model:

The DT-CNN hybrid model outperforms the individual models, achieving near-perfect accuracy, precision, recall, and F1-score. This model effectively combines the interpretability of Decision Tree with the pattern recognition capabilities of Convolutional Neural Network. The Decision Tree provides an initial classification, excelling in clear cases of fraud and non-fraud. The CNN then focuses on refining the misclassifications from the Decision Tree, leveraging its ability to detect intricate patterns. This hybrid approach addresses the limitations of both models, offering high interpretability and robust performance, especially in handling imbalanced datasets.

The DT-CNN model's remarkable performance demonstrates its efficacy in minimizing both false positives and false negatives, making it a superior choice for fraud detection. The hybrid model's ability to enhance precision and recall significantly reduces the financial risks associated with misclassification, providing a comprehensive and reliable solution for credit card fraud detection.

VI. CONCLUSION

The comparative analysis highlights the strengths and weaknesses of each model. The DT-CNN hybrid model consistently outperforms individual Decision Tree and CNN models by leveraging the strengths of both techniques to optimize accuracy, precision, recall, and F1 score. Despite the challenges posed by an imbalanced dataset, this combined approach proves to be a robust solution for fraud detection.

The DT-CNN hybrid model's significance extends beyond credit card fraud detection, offering potential advancements in accuracy and reliability across various critical applications. In medical diagnosis, misclassification can lead to incorrect treatment plans, while in manufacturing, it can result in defective products reaching consumers. In environmental monitoring and disaster prediction, misclassification of early warning signs can have devastating consequences, including loss of life and property damage.

By enhancing the accuracy of environmental data analysis, the DT-CNN hybrid model can improve disaster management efforts, mitigate potential risks, and reduce the likelihood of catastrophic events. This model's adaptability and stellar performance make it a versatile tool across diverse domains, poised to enhance operational efficiency, reduce risks, and improve decision-making processes.

Future work could explore further enhancements, such as data augmentation or ensemble methods, to improve the detection of fraudulent transactions and expand the model's applications. By doing so, the DT-CNN hybrid model can have a profound impact on saving lives, preventing property damage, and promoting a safer, more reliable future.

REFERENCES

[1] Mullen, Caitlin. “Card Industry’s Fraud-Fighting Efforts Pay Off: Nilson Report.” *Payments Dive*, January 5, 2023. https://www.paymentsdive.com/news/card-industry-fraud-fighting-efforts-pay-off-nilson-report-credit-debit/639675 .

[2] Takyar, Akash. “AI in Fraud Detection: Use Cases, Benefits, Solution and Implementation.” *LeewayHertz*, May 27, 2024. https://www.leewayhertz.com/ai-in-fraud-detection/

[3] O’Connor, Ade. “North American Ecommerce and Retail Companies Face a $3.00 Total Cost for Each Dollar Lost to Fraud, According to True Cost of Fraud Study from LexisNexis® Risk Solutions.” *LexisNexis Risk Solutions*, March 27, 2024.

https://risk.lexisnexis.com/about-us/press-room/press-release/20240327-tcof-retail-ecommerce

[4] Liu, Henry, and Staff in the Office of Technology. "New Data Shows FTC Received 2.8 Million Fraud Reports from Consumers in 2021." *Federal Trade Commission*, February 22, 2022. https://www.ftc.gov/news-events/news/press-releases/2022/02/new-data-shows-ftc-received-28-million-fraud-reports-consumers-2021-0

[5] Ali, Najwan, Shahad Hasan, Ahmad Ghandour, and Zainab Al-Hchimy. "Improving Credit Card Fraud Detection Using Machine Learning and GAN Technology." *BIO Web of Conferences* 97 (2024): 00076. https://doi.org/10.1051/bioconf/20249700076

[6] Mijwil, M. M., and I. E. Salem. "Credit Card Fraud Detection in Payment Using Machine Learning Classifiers." *Asian Journal of Computer and Information Systems* 8, no. 4 (2020): 50. Asian Online Journals. http://www.ajouronline.com

[7] Ileberi, E., Y. Sun, and Z. Wang. "A Machine Learning Based Credit Card Fraud Detection Using the GA Algorithm for Feature Selection." *Journal of Big Data* 9 (2022): 24. https://doi.org/10.1186/s40537-022-00573-8

[8] Asha, R., S. K.-G. T. Proceedings, and undefined. "Credit Card Fraud Detection Using Artificial Neural Network." *Elsevier*. Accessed March 11, 2023. https://www.sciencedirect.com/science/article/pii/S2666285X21000066

[9] Chang, Victor, Le Minh Thao Doan, Alessandro Di Stefano, Zhili Sun, and Giancarlo Fortino. "Digital Payment Fraud Detection Methods in Digital Ages and Industry 4.0." *Computers & Electrical Engineering* 100 (2022): 107734.

https://doi.org/10.1016/j.compeleceng.2022.107734

[10] Sadineni, Praveen Kumar. "Detection of Fraudulent Transactions in Credit Card Using Machine Learning Algorithms." In *Proceedings of the International Conference on ISMAC (IoT in Social, Mobile, Analytics and Cloud)*, 659-660. 2020.

https://doi.org/10.1109/I-SMAC49090.2020.9243545

[11] [11] Błaszczyński, J., A. T. de Almeida Filho, A. Matuszyk, M. Szeląg, and R. Słowiński. "Auto Loan Fraud Detection Using Dominance-Based Rough Set Approach Versus Machine Learning Methods." *Expert Systems with Applications* 163 (2021): 113740. https://doi.org/10.1016/j.eswa.2020.113740

[12] Madhavi, M., et al. "Credit Card Fraud Detection Using CNN." 2023. https://www.ijrti.org/papers/IJRTI2304141.pdf

APPENDIX A

Term	Definition
Accuracy	This metric represents the percentage of correct predictions made by the model. accuracy = $\frac{TP+TN}{TP+TN+FN+FP}$
Binary Cross Entropy	A loss function is used for binary classification problems that measures the performance of a model whose output is a probability value between 0 and 1. It calculates the difference between the actual class and the predicted probability.
Confusion Matrix	A confusion matrix is a tabular representation that shows the actual versus predicted classifications made by the model. It helps in visualizing the performance of a classification algorithm.
Conv1D	Conv1D refers to a one-dimensional convolutional layer.
Criterion - entropy	Entropy helps to determine the best split by selecting the attribute that results in the most significant information gain.
Dense	A fully connected layer in a neural network where each neuron receives input from all neurons of the previous layer, used for learning complex representations of the input data.
F1 Score	The F1 score represents the harmonic mean of precision and recall, providing a balanced measure of model performance. F1 score = $2 \times \frac{Precision \times Recall}{Precision + Recall}$
MaxPooling1D	A down-sampling operation that reduces the dimensionality of the input, typically used in CNNs.
Precision	Precision refers to the percentage of positive predictions that are correct. precision = $\frac{TP}{TP+FP}$
Recall	Recall is the fraction of actual positives that are correctly predicted by the model. recall = $\frac{TP}{TP+FN}$
RMSProp Optimizer	Root Mean Square Propagation is an optimization algorithm that is designed to adjust the learning rate of each parameter individually, based on the average of recent magnitudes of the gradients for that parameter.
SMOTE	Synthetic Minority Oversampling Technique, is a machine learning technique that balances class distribution in datasets with imbalanced data.

Addressing Water and Sanitation Challenges in Rural Afghanistan: Barriers, Initiatives, and Sustainable Solutions

Sayed Basir Ahmad AYOUBI[1*], Gulam Hassan Haidary[2]

[1]Department of Civil Engineering, Jami University, Herat, Afghanistan
ORCID: https://orcid.org/ 0000-0002-6110-2276
[2]Department of Civil Engineering, Jami University, Herat, Afghanistan
ORCID https://orcid.org/0009-0001-7344-1021
*Corresponding Author

Received: 05 July 2024/ Revised: 10 July 2024/ Accepted: 18 July 2024/ Published: 31-07-2024

Abstract— *This comprehensive study addresses the critical challenges of providing clean water and proper sanitation in rural Afghanistan. Despite significant international aid and various governmental initiatives, rural communities face persistent issues related to water scarcity, contaminated sources, and inadequate sanitation facilities. These challenges are compounded by political instability, cultural practices, and economic constraints. The hypothesis posits that integrated, community-driven, and sustainable interventions are essential to overcome these obstacles. The objectives include identifying key barriers, evaluating current initiatives, and proposing viable strategies for improvement. A thorough literature review underpins the study, highlighting the severity and complexity of the crisis. Research methods incorporate qualitative approaches such as interviews, focus groups, and observational studies. Data analysis focuses on thematic patterns and practical gaps. The conclusion emphasizes the need for coordinated efforts and continuous monitoring to ensure long-term success.*

Keywords— ***Sustainable development, Clean water, Water scarcity, and Rural Afghanistan.***

I. INTRODUCTION

Access to clean water and proper sanitation is fundamental to human health and development. In rural Afghanistan, however, these basic necessities remain largely unmet. The lack of clean water and sanitation exacerbates health issues, economic hardship, and social inequalities. This article aims to explore the multifaceted challenges faced in providing clean water and proper sanitation in rural Afghanistan. By understanding these challenges and examining potential solutions, we hope to contribute to the improvement of living conditions in these communities.

The article is structured to provide a detailed exploration of the subject. It begins with a hypothesis, followed by clearly defined objectives. A thorough literature review sets the context, while the research methods and data analysis sections detail the approach and findings. The conclusion offers actionable recommendations and highlights the importance of sustained efforts in addressing these issues.

1.1 Hypothesis:

The central hypothesis of this study is that targeted, community-based interventions and sustainable infrastructure development are crucial for improving water and sanitation in rural Afghanistan. Additionally, it posits that a holistic approach, integrating local knowledge, international aid, and government policy, is necessary to overcome these challenges effectively.

1.2 Objectives:

1. Identify the primary barriers to clean water and sanitation in rural Afghanistan.
2. Assess the effectiveness of current water and sanitation initiatives.
3. Propose sustainable and community-driven solutions to enhance water and sanitation services.
4. Highlight the role of international aid and government policies in addressing these challenges.

5. Provide recommendations for future research and interventions.

II. LITERATURE REVIEW

2.1 Water and Sanitation in Rural Afghanistan: An Overview:

In rural Afghanistan, only a small fraction of the population has access to improved water sources and sanitation facilities. According to UNICEF, merely 27% of rural Afghans have access to safe drinking water, and only 20% have proper sanitation. The lack of these basic services leads to widespread health problems, including waterborne diseases such as diarrhea, which is a major cause of child mortality.

2.2 Health Impacts:

The World Health Organization (WHO) emphasizes the critical link between inadequate water and sanitation and the prevalence of diseases. Waterborne diseases, including cholera, dysentery, and typhoid, are rampant due to the consumption of contaminated water. Poor sanitation practices further exacerbate these health issues, leading to frequent outbreaks of diseases.

2.3 Socio-economic Consequences:

The Afghanistan Research and Evaluation Unit (AREU) and various non-governmental organizations (NGOs) highlight the socio-economic impacts of poor water and sanitation. Women and children are particularly affected, as they are often responsible for fetching water, a task that can take several hours each day. This time-consuming chore limits educational and economic opportunities, perpetuating cycles of poverty.

2.4 Political and Infrastructural Challenges:

Political instability and ongoing conflict have severely hampered efforts to improve water and sanitation infrastructure. Government initiatives are often underfunded and poorly executed due to corruption and lack of coordination. The challenging terrain and dispersed populations in rural areas further complicate infrastructure development.

2.5 Cultural and Behavioral Factors:

Cultural practices and beliefs also play a significant role in water and sanitation issues. For example, open defecation remains prevalent in many rural communities due to a lack of awareness about its health impacts and social norms. Changing these behaviors requires sustained education and community engagement.

2.6 International Aid and NGO Interventions:

Various international organizations, including UNICEF, WHO, and numerous NGOs, have been actively involved in water and sanitation projects in Afghanistan. These efforts have led to some improvements, but many challenges remain. Successful projects often involve community participation and local ownership, highlighting the importance of culturally sensitive and sustainable approaches.

2.7 Sustainable Development Goals (SDGs):

The United Nations' Sustainable Development Goals, particularly Goal 6, which aims to ensure availability and sustainable management of water and sanitation for all, provide a framework for addressing these issues. Achieving this goal in Afghanistan requires a concerted effort from all stakeholders, including the government, international organizations, and local communities.

III. RESEARCH METHODS

This study employs a qualitative approach to explore the challenges and potential solutions for providing clean water and proper sanitation in rural Afghanistan. The methods include:

3.1 Interviews:

Structured interviews were conducted with a diverse group of stakeholders, including local residents, government officials, and representatives from NGOs. These interviews provided insights into the on-the-ground realities and the perspectives of different stakeholders.

3.2 Focus Groups:

Focus groups were held with community members to facilitate in-depth discussions about their needs, challenges, and suggestions for improvement. This method helped gather a wide range of views and fostered a collaborative atmosphere for generating essential to ensure that interventions remain effective and responsive to changing conditions.

IV. RECOMMENDATIONS

Based on the findings, several recommendations can be made to improve the provision of clean water and proper sanitation in rural Afghanistan:

1. **Strengthen Community Involvement:**
   - Foster community ownership of water and sanitation projects to ensure sustainability.
   - Involve community members in the decision-making process to ensure solutions are culturally appropriate and locally relevant.
2. **Enhance Education and Awareness:**
   - Implement comprehensive hygiene education programs in schools and communities.
   - Conduct awareness campaigns to highlight the health benefits of proper sanitation and clean water practices.
3. **Promote Sustainable Technologies:**
   - Invest in sustainable and low-maintenance technologies, such as solar-powered water pumps and eco-friendly latrines.
   - Encourage the use of local materials and labor to build and maintain water and sanitation infrastructure.
4. **Improve Coordination and Collaboration:**
   - Enhance coordination between government agencies, NGOs, and international donors to avoid duplication of efforts and ensure efficient use of resources.
   - Establish multi-stakeholder platforms to facilitate information sharing and collaboration.
5. **Increase Funding and Investment:**
   - Advocate for increased funding from international donors and the Afghan government to support water and sanitation projects.
   - Explore innovative financing mechanisms, such as microfinance and public-private partnerships, to mobilize additional resources.
6. **Monitor and Evaluate:**
   - Develop robust monitoring and evaluation frameworks to track the progress and impact of water and sanitation initiatives.
   - Use data and feedback to adapt and improve interventions continuously.
7. **Support Policy Development:**
   - Advocate for the development and enforcement of policies and regulations that support sustainable water and sanitation practices.
   - Promote accountability and transparency in the management of water and sanitation resources.
8. **Address Gender and Social Inequities:**
   - Ensure that water and sanitation projects address the specific needs of women, children, and marginalized groups.
   - Promote gender-sensitive approaches and involve women in the planning and implementation of projects.

V. DETAILED SECTIONS FOR EACH OBJECTIVE

5.1 Objective 1: Identify the Primary Barriers to Clean Water and Sanitation in Rural Afghanistan

Geographical Challenges: The rugged and mountainous terrain of Afghanistan poses significant logistical challenges for the construction and maintenance of water and sanitation infrastructure. Remote villages are often inaccessible, especially during harsh weather conditions, which hinders the delivery of materials and services.

Political Instability: Years of conflict and political instability have disrupted development efforts. Infrastructure projects are frequently delayed or abandoned due to security concerns, and existing facilities are often damaged or destroyed during conflicts.

Economic Constraints: High levels of poverty in rural Afghanistan limit the ability of households to invest in improved water and sanitation facilities. Limited economic opportunities also reduce the capacity of communities to maintain and repair infrastructure.

Cultural Practices: Traditional beliefs and practices, such as open defecation and the use of contaminated water sources, are deeply ingrained in some rural communities. These practices are often perpetuated by a lack of awareness about their health impacts.

Lack of Awareness: Many rural residents are unaware of the health risks associated with poor water and sanitation practices. This lack of awareness hampers efforts to promote behavior change and adopt improved practices.

5.2 Objective 2: Assess the Effectiveness of Current Water and Sanitation Initiatives

Successful Projects: Successful water and sanitation projects in rural Afghanistan often share common characteristics, including strong community involvement, local ownership, and integration with other development efforts. These projects typically include components such as hygiene education, capacity building, and the use of sustainable technologies.

Challenges Faced: Despite some successes, many initiatives face significant challenges. These include insufficient funding, poor coordination among stakeholders, cultural resistance, and logistical difficulties. Projects that do not adequately consider local contexts and community needs often fail to achieve their objectives.

5.3 Objective 3: Propose Sustainable and Community-Driven Solutions

Community Engagement: Community engagement is crucial for the success of water and sanitation projects. Involving community members in planning, implementation, and maintenance ensures that solutions are locally appropriate and sustainable. Community-based organizations can play a key role in mobilizing resources and promoting behavior change.

Education and Training: Continuous education and training on hygiene practices and the maintenance of facilities are essential for sustaining improvements. Schools, community centers, and local media can be used to disseminate information and raise awareness.

Integrated Approaches: Integrating water and sanitation initiatives with other development efforts, such as health, education, and livelihood programs, can create synergies and enhance overall impact. For example, combining hygiene education with school programs can promote healthier behaviors among children and their families.

Sustainable Technologies: Investing in sustainable and low-maintenance technologies, such as solar-powered water pumps, gravity-fed water systems, and eco-friendly latrines, can improve the feasibility and longevity of water and sanitation infrastructure. Using local materials and labor can also reduce costs and ensure that communities have the skills needed to maintain facilities.

5.4 Objective 4: Highlight the Role of International Aid and Government Policies

International Aid: International aid has played a crucial role in funding and supporting water and sanitation projects in Afghanistan. However, to be effective, aid must be aligned with local needs and priorities. Donors should work closely with local communities and governments to ensure that projects are sustainable and culturally appropriate.

Government Policies: The Afghan government needs to strengthen its policies and regulations related to water and sanitation. This includes increasing funding, improving coordination among agencies, and promoting accountability. Developing and enforcing standards for water quality and sanitation facilities is also essential.

5.5 Objective 5: Provide Recommendations for Future Research and Interventions

Innovative Solutions: Future research should focus on developing and testing innovative solutions to water and sanitation challenges. This could include exploring new technologies, financing mechanisms, and community engagement strategies.

Impact Evaluation: Evaluating the impact of current initiatives is crucial for understanding what works and why. Future interventions should be based on evidence from rigorous impact evaluations.

Enhanced Coordination: Improving coordination among stakeholders, including government agencies, NGOs, and international donors, is essential for avoiding duplication of efforts and ensuring efficient use of resources. Establishing multi-stakeholder platforms can facilitate information sharing and collaboration.

VI. APPENDICES

6.1 Appendix 1: Interview Questions

A sample of interview questions used in the study:

1. What are the main challenges you face in accessing clean water?
2. How has the lack of proper sanitation affected your community?
3. What initiatives have been undertaken to improve water and sanitation in your area?
4. How involved is the community in these initiatives?
5. What do you think could be done to improve water and sanitation services?

6.2 Appendix 2: Focus Group Discussion Guide

A sample of topics covered in focus group discussions:

1. Perceptions of water quality and sanitation practices.
2. Barriers to accessing clean water and proper sanitation.
3. Community participation in water and sanitation projects.
4. Impact of water and sanitation issues on health and livelihoods.
5. Suggestions for improving water and sanitation services.

Appendix 3: Observational Study Checklist

A checklist used for observational studies:

1. Condition of water sources (e.g., wells, rivers).
2. Functionality of water supply systems (e.g., pumps, pipes).
3. Condition of sanitation facilities (e.g., latrines, sewage systems).
4. Hygiene practices observed in the community.
5. Maintenance and repair activities undertaken by the community.

ACKNOWLEDGEMENTS

We thank Jami University for their continuous support and encouragement throughout this research journey. The faculty and staff provided invaluable resources and guidance that were essential to the completion of this study. We also extend our heartfelt thanks to the community members of rural Afghanistan who participated in our interviews and focus groups, sharing their experiences and insights with us. Your cooperation and openness were invaluable to this study. Additionally, we acknowledge the funding and support provided by Jami University. Without their contributions, this research would not have been possible.

REFERENCES

[1] Afghanistan Research and Evaluation Unit. (2019). Water Management in Rural Afghanistan.

[2] Ahmad, S., & Khan, A. (2019). Cultural practices and water sanitation in Afghanistan. Health and Hygiene Journal.

[3] BRAC. (2020). Community-based Water Management in Afghanistan.

[4] CARE. (2021). Improving Water and Sanitation in Afghanistan.

[5] Global Water Partnership. (2021). Integrated Water Resources Management in Afghanistan.

[6] International Rescue Committee. (2019). Water and Sanitation Services in Rural Afghanistan.

[7] Jha, A. K., et al. (2018). Infrastructure for sustainable development in Afghanistan. International Journal of Water Resources Development.

[8] Jones, T., & Shah, R. (2022). Sustainable water management in conflict zones. International Journal of Water Governance.

[9] Mercy Corps. (2019). Rural Water Supply Systems in Afghanistan.

[10] National Rural Water Association. (2018). Rural Water Challenges in Afghanistan.

[11] Norwegian Refugee Council. (2021). Water, Sanitation, and Hygiene in Afghanistan.

[12] Oxfam. (2020). Water Security in Rural Afghanistan.

[13] Plan International. (2021). Sanitation Challenges in Afghan Rural Areas.

[14] Save the Children. (2020). Health and Hygiene in Rural Afghan Schools.

[15] Smith, L., et al. (2020). The impact of war on water and sanitation infrastructure in Afghanistan. Journal of Conflict and Health.

[16] Tearfund. (2020). Sustainable Water Solutions in Afghanistan.

[17] Thompson, P., & Harris, M. (2021). Community-driven approaches to water and sanitation in rural Afghanistan. Water Policy Journal.

[18] UNICEF. (2020). Water, Sanitation and Hygiene.

[19] United Nations. (2022). Sustainable Development Goals: Clean Water and Sanitation.

[20] United States Agency for International Development. (2019). Afghanistan Water and Sanitation Program.

[21] WaterAid. (2019). Ensuring Safe Water and Sanitation in Afghanistan.

[22] World Bank. (2020). Afghanistan: Water Supply and Sanitation.

[23] World Health Organization. (2021). Sanitation and Health.

[24] World Vision. (2020). Access to Clean Water in Afghanistan.

AD Publications
Sector-3, MP Nagar, Bikaner,
Rajasthan, India
www.adpublications.org, info@adpublications.org