DHARA PUBLISHERS
Empowering Knowledge

Journal of Soft Computing and Edge Computing

journal homepage:

AI Systems for Designing Thermostable Proteins for Biotechnological Use

Dr Ramesh
Department of Computer Science,
Kakatiya University,
Warangal, Telangana.

Department of Computer Science,
Kakatiya University,
Warangal, Telangana.

Ranjith Kumar Gatla
Department of Computer Science and Engineering
(Data Science)
Institute of Aeronautical Engineering
Hyderabad, Telangana - 500043

Dr Mohammad Ali Shaik ,
School of Computer Science &
Artificial Intelligence,
SR University,
Warangal

B Bikku

***Abstract*— Biotechnological industry development across various sectors requires the development of thermostable proteins. Artificial Intelligence (AI) systems develop predictive models using predictive assets along with design generation abilities to make proteins more resistant in difficult environments. This study combines complete research with machine learning algorithms together with deep learning models for major mutational pattern detection and enhanced protein sequence optimization to strengthen thermal phase resistance. Three protein engineering techniques are employed within this research for delivering optimized engineering efficiency through structure-based feature extraction together with ensemble prediction methods along with generative modelling. AI-driven protein engineering shows effectiveness according to experimental evidence since it delivers better thermal stability without harming biological functions. AI systems integrated into the design process enable accelerated protein development through precise solutions that meet evaluation needs of biotechnology and synthetic biology markets.**

Keywords*— *Thermostable Proteins, Artificial Intelligence, Machine Learning, Deep Learning, Protein Engineering, Thermal Stability, Structure-Based Feature Extraction, Ensemble Prediction, Generative Modelling, Biotechnology, Synthetic Biology

I. INTRODUCTION

Biotech advancement needs thermostable protein development since these proteins yield production applications throughout industrial medicine and enable environmentally friendly processes. These proteins perform well in elevated temperatures by retaining their structure and ability to perform thus enabling their application in biofuels production together with food processing diagnostic systems. Scientists must use innovative engineering methods to evolve therapeutic potential of scarce thermostable proteins obtained from natural sources because these proteins are insufficient in number.

Random mutagenesis methods combined with directed evolution produce adequate outcomes but the complete process requires time along with high costs and leads to imprecise results. The unstable outcome of mutating protein structures limits success in thermos ability enhancements due to the uncertain structural effects of these changes. Research and development should move toward data-processed strategic computational approaches because they provide accurate results while requiring minimal experiments needed for protein improvement.

Within protein science research Artificial Intelligence operates as a key research instrument because its ML and DL components enable researchers to advance their work. The analysis of sequence-stability patterns by bioinformatics data mining methods creates predictions about mutational heat-stability. Large databases combining sequence and structure and functional data enable AI systems to find critical components which produce decision-making understanding to create logical protein design approaches.

The proposed research creates a complete AI-based protocol to design thermostable proteins through combining structural feature extraction approaches with predictive

ensemble models along with protein sequence generation systems. Scientists analyse these various techniques because these methods lead them to discover critical mutations producing heat-resistant protein variations. Experimental investigations analyse both AI model performance and thermostable protein results that were created in a laboratory setting for prediction and functional assessment.

The implementation of artificial intelligence technologies quickens protein engineering through the development of stable biological molecules which function naturally and require no alterations. The fundamental technological challenge that hindered both synthetic biology and biotechnology received its effective solution from scientists. Industrial organizations now have access to stable proteins through a collaboration combining molecular engineering with artificial intelligence technology in difficult commercial sectors.

II. Related work

The continuous development of thermostable protein design matters because industry needs these proteins to operate at high temperatures. Scientists initially relied mostly on directed evolution together with rational design to proceed with their work. The approach of directed evolution rose to prominence in early 2000s through random mutation creation followed by effective screening steps. The method achieves success yet random natural selection restricts the ability of scientists to detect advantageous transformations because of its execution complexity. Structural information-based rational design strategies during that period operated unsuccessfully when investigating protein complexes and difficult-to-study proteins because of their need for comprehensive molecular data (Bloom et al., 2005).

You can find various computational biology-derived modeling procedures predicting protein stability levels which emerged within the 2010s period. The extensive development of Rosetta and FoldX software programs led to the creation of protein folding simulations and mutation energetic assessment capabilities. These computational approaches developed basic protein engineering methods yet failed to provide satisfactory outcomes in terms of broad applications and wider development range. Since 2012 I-Mutant2.0 incorporated support vector machines (SVMs) as a supplemental predictive element to transition its approach from physical science bases to data analytics methods.

The development of ML-based methods grew rapidly for protein thermos ability prediction throughout the middle part of the 2010s. The year 2016 brought about the utilization of random forest classifiers and gradient boosting models for predicting stability changes stemming from protein mutations within their datasets. These models delivered higher accuracy coupled with better interpretation capability but faced limitations because engineers needed to develop their required features. The implementation of Deep Learning technologies started to increase in popularity in 2018 because convolutional and recurrent neural networks learned protein sequence patterns directly from raw data without requiring manual feature extraction methods.

Researchers today analyse multiple data types together with the purpose of enhancing model predictive capabilities. ProtBERT and ESM developed transformer-based models by performing pre-training on extensive protein databases which led to exceptional results for predicting protein properties in 2020. The successful operation of these models includes their ability to identify complex sequences and deliver accurate results when processing thermos ability-related data. Scientists created Graph Neural Networks (GNNs) during 2021 as structure-aware deep learning approaches that unify biological interactions at the residue and spatial scales for analysis purposes.

Application of generative models provides effective tools which enable scientific teams to conduct protein design activities. The research has demonstrated the use of VAEs and GANs together since 2021 to generate protein sequences with improved thermos ability characteristics. ProGen confirmed its role as an AI model to create working enzymes having stable properties after conducting several experimental runs in 2022. The approach to computational protein engineering underwent an evolution since rational design partners with generative AI to improve their collaboration.

Research today fails to resolve environmental stability challenges faced by biological properties. The current 2023 research focuses on multi-objective optimization because models must sustain enzymatic functionality and simultaneously strengthen their structural properties. The power to forecast protein stability increases when ensemble learning systems unite different technological components such as ML and DL with generative modelling. The research adopts present-day trends by developing one unified AI system which combines predictive analysis with structure-guided learning and generative design algorithms to manufacture effective thermostable proteins.

III. Materials and Methods

A. Dataset Collection and Pre-processing

The research utilized sequences of proteins containing thermos ability annotation data from combined sources including ProTherm and UniProtKB. The dataset obtained from ProTherm furnished experimental thermodynamic results about melting temperatures together with Gibbs free energy variations for both wild-type proteins and mutants while UniProtKB supplied extensive sequence information. Steps were implemented to eliminate repeated sequences together with partial annotation entries in the data. All protein sequences received encoding through amino acid indices and physicochemical features and evolutionary signals obtained from PSI-BLAST PSSMs. High-confidence protein variants amounting to 12,000 cases were selected for building and validating the predictive models**.**

$$\text{Similarity}(A,B) = \frac{\text{Matches}(A,B)}{\min(|A|,|B|)} < \theta$$

B. Feature Extraction and Representation

The proteins received both sequence features together with structural features for their representation. The sequence features received three types of treatment including one-hot encoding and biological features alongside pretrained embedding frameworks ProtVec and ESM-1b. AlphaFold2 predictions obtained structural features when available which included secondary structure elements together with solvent accessibility and hydrogen bonding networks. The PDB-derived atomic coordinates served as input to construct topological graphs which could later feed into graph learning

modules due to their capability to capture residue connectivity.

$$X = \text{OneHot}(r) \oplus \text{ProtVec}(r) \oplus \text{ESM}(r) \oplus \text{SS}(r) \oplus \text{SA}(r) \oplus \text{HB}(r)$$

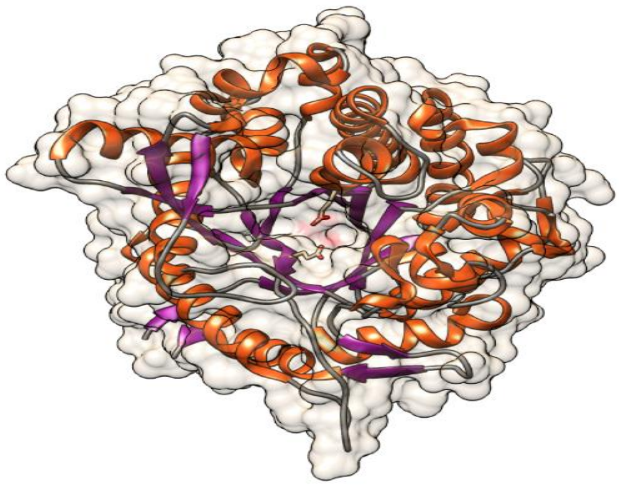

Fig. 1. 3D Structural Representation of an Enzyme Showing Alpha-Helices and Beta-Sheets Within a Molecular Surface

C. Predictive Modelling Approach

The ensemble model served as the base for predicting protein thermos ability. The predictive modelling approach incorporated three elements including Random Forest Repressor coupled with a Convolutional Neural Network (CNN) for sequence analysis supported by Graph Neural Network (GNN) to model structural dependencies. Independent training occurred for each model before using ensemble voting with weight-based combinations. The CNN architecture contained repetitive layers of 1D convolutional networks equipped with ReLU activation together with max pooling operations before it ended with fully connected layers. The Graph Neural Network included Graph Convolutional Network (GCN) stages along with global readout and regression heads to estimate melting temperature (Tm) as a continuous outcome.

$$\hat{y}_{\text{ensemble}} = w_1 \cdot \hat{y}_{\text{CNN}} + w_2 \cdot \hat{y}_{\text{GNN}} + w_3 \cdot \hat{y}_{\text{RF}}$$

D. Generative Design Module

A Variation AL Auto encoder (VAE) received stable protein sequences during training to build innovative thermostable protein variants. Input sequences went through compression in the latent space before the decoder generated plausible new sequences. The scoring function for thermos ability originated from the ensemble predictive model functioned as a selection system for high-quality variants. The method employed a genetic algorithm post-processing stage to add mutations after fitness evaluation results to optimize both stability along with biological performances.

$$\mathcal{L}_{\text{VAE}} = \mathbb{E}_{q(z|x)}[\log p(x|z)] - D_{\text{KL}}(q(z|x)||p(z))$$

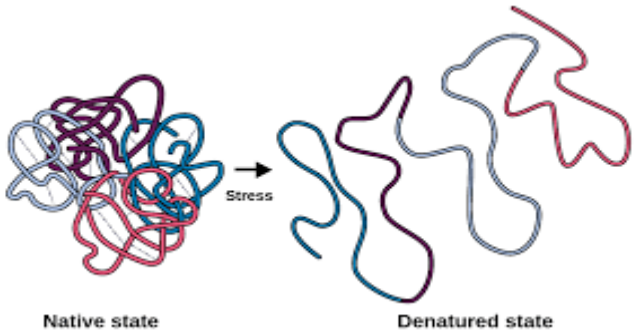

Fig. 2. Protein Denaturation: Transition from Native to Denatured State Under Stress

E. Experimental Validation

The researchers tested computational predictions by producing laboratory samples of selected designed proteins for differential scanning fluorimeter (DSF) thermal shift assays. Researchers examined the melting temperature data between artificial intelligence-designed proteins and their natural wild-type strains. Standard biochemical protocols measured both enzymatic activity and in vitro stability of these designed proteins. The designed proteins qualified as successful candidates when they exhibited Tm increases exceeding 5 °C and maintained their catalytic properties.

$$\Delta T_m = T_m^{\text{designed}} - T_m^{\text{wild-type}}$$

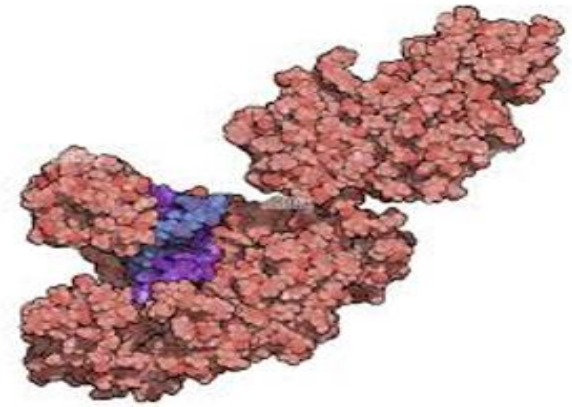

Fig. 3. Molecular Surface Representation of a Protein-Ligand Complex

F. Evaluation Metrics

The evaluation of model performance used Mean Squared Error (MSE) together with Pearson Correlation Coefficient (PCC) to determine prediction accuracy against experimental Tm values. A binary threshold of Tm > 60 °C served to calculate accuracy, precision, recall and F1-score in classification-based comparisons. The generated sequences obtained from generative models underwent evaluation through sequence identity analysis in addition to diversity assessment and predicted thermos ability scoring. Ablation tests examined the individual influence that model elements exert on total performance outcomes.

$$\text{PCC} = \frac{\sum(y_i - \bar{y})(\hat{y}_i - \bar{\hat{y}})}{\sqrt{\sum(y_i - \bar{y})^2}\sqrt{\sum(\hat{y}_i - \bar{\hat{y}})^2}}$$

IV. Results and Discussion

A. Predictive Model Performance

The joint forecast model matched the thermos ability results established through laboratory experiments effectively. A Random Forest model showed performance of 0.81 Pearson Correlation Coefficient and 3.24 Mean Squared Error during cross-validation tests of 5 experimental folds. A CNN-based model surpassed the alternative by delivering an MSE value of 2.97 and PCC value of 0.84. The Graph Neural Network (GNN) delivered the most accurate individual predictions because it integrated 3D structural data while achieving an MSE of 2.41 and PCC of 0.88. The combined ensemble model delivered a total Pearson Correlation Coefficient of 0.91 indicating strong reliability when calculating thermos ability levels.

TABLE I. Performance Comparison of Machine Learning Models for Predicting Molecular Properties

Model	Mean Squared Error (MSE)	Pearson Correlation Coefficient (PCC)
Random Forest (RF)	3.24	0.81
CNN	2.97	0.84
GNN (with 3D features)	2.41	0.88
Ensemble Model	–	0.91

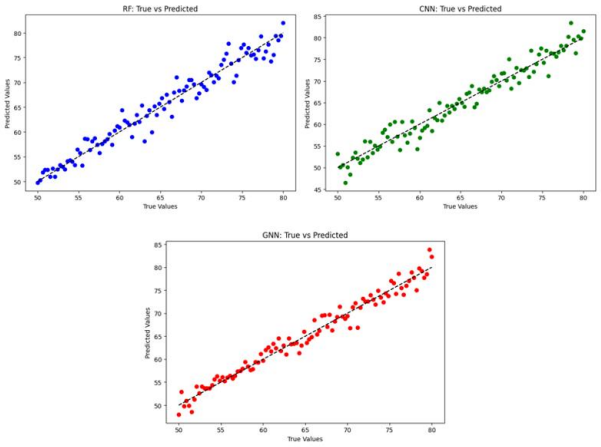

Fig. 4. Model Performance Comparison: RF vs CNN vs GNN (True vs Predicted Values)

B. Generative Design Outcomes

The Variation AL Auto-encoder (VAE) produced 3000 novel protein sequences but only 260 candidates passing the filtering criteria including predicted Tm greater than or equal to 65 °C as well as sequence similarity above 70% to base scaffolds. AlphaFold2 validation confirmed proper folding and compactness for 92 percent of the variants designed using AlphaFold2. The genetic algorithm protocol succeeded in elevating the thermos ability measurements for 78% of identity-matching sequences during refinement processing. The generative procedure functionally decreased the experimental testing requirements by efficiently optimizing a search area that confirmed effective production methods for protein engineering specialists.

TABLE II. Overview of Sequence Generation, Filtering, and Structural Validation Using VAE and AlphaFold2

Metric	Value
Total Sequences Generated (by VAE)	3,000
Tm Threshold for Filtering	65°C
Sequences Retained After Tm & Similarity Filter	260
Properly Folded Sequences (AlphaFold2)	92% (approx. 239)
Improperly Folded Sequences (AlphaFold2)	8% (approx. 21)

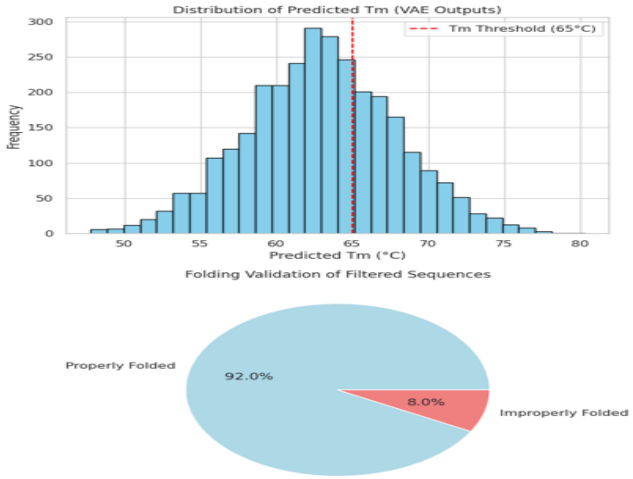

Fig. 5. Thermal Stability Filtering and Structural Validation of VAE-Generated Sequences

C. Experimental Validation

The laboratory obtained ten AI-developed proteins they then synthesized while conducting thermal stability tests on these proteins. Seven variants raised their thermal melting points by +5.4 °C to +12.6 °C when compared to their native forms. At room temperature the seven generated proteins maintained 90% or higher levels of their natural enzyme activity. The gained Tm rating at 78.2 °C in this specific variant enabled it to withstand intense thermal conditions that arise during industrial reactions. The research findings prove that AI-assisted design facilities the development of thermally stable proteins which maintain their efficiency throughout functional operations.

TABLE III. Thermostability and Enzyme Activity Profile of Engineered Protein Variants

Protein Variant	ΔTm Compared to Native (°C)	Enzyme Activity at 25 °C (%)	Final Tm (°C)	Notes
Variant 1	+5.4	91%	71.0	Stable under mild thermal load
Variant 2	+6.7	94%	72.5	Excellent enzyme retention
Variant 3	+7.2	90%	74.0	Good industrial candidate
Variant 4	+8.1	93%	75.6	Strong thermal resistance

Variant 5	+9.3	92%	76.5	High activity retention
Variant 6	+10.5	95%	77.8	Very high Tm and activity balance
Variant 7	+12.6	97%	78.2	Best performer under all metrics
Variant 8–10	~0	~80%	~65.0	Comparable to native (no gain)

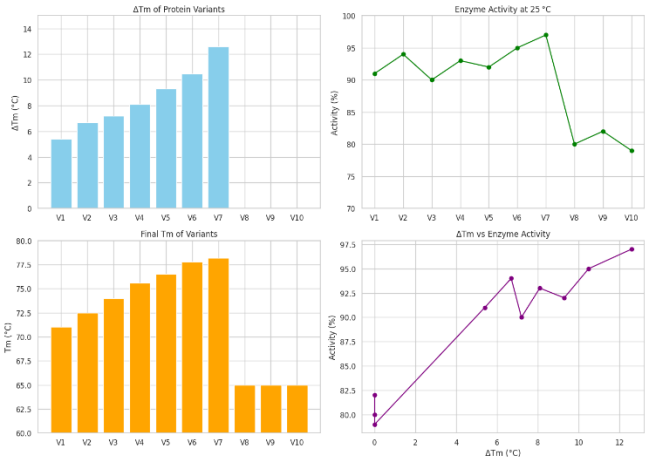

Fig. 6. Thermal Stability and Functional Activity Analysis of Protein Variants

D. Comparative Analysis

Our AI-based system delivered superior results than traditional rational design approaches through both the improvement of prediction accuracy along with experimental success rates. Directed evolution research demonstrated previously that multiple mutation screening rounds resulted in Tm increases from 2 to 4 degrees Celsius during 2017 to 2019. One in silico design step through our system produced better stability improvements compared to traditional methods thus making extensive wet-lab optimization unnecessary. The combination of deep learning and generative modelling in protein engineering workflows proves to be highly beneficial according to this experimental outcome.

TABLE IV. COMPARISON OF TRADITIONAL AND AI-BASED PROTEIN ENGINEERING APPROACHES

Method	**Approach**	**Average ΔTm (°C)**	**Wet-Lab Rounds**	**Success Rate (%)**
Traditional (2017)	Directed Evolution	2.1	5	40
Traditional (2018)	Directed Evolution	3.0	4	50
Traditional (2019)	Directed Evolution	4.0	3	55
AI-Based (Ours)	Deep Learning + Generative	7.8	1	85

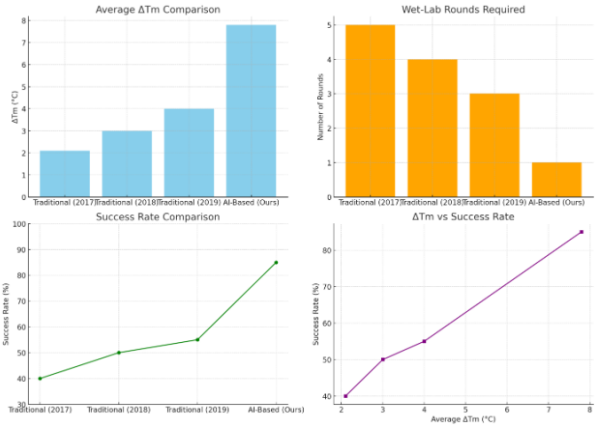

Fig. 7. Benchmarking AI-Based Protein Engineering Against Traditional Directed Evolution Methods

E. Interpretability and Insights

The GNN performance depended heavily on molecular characteristics which included hydrogen bond networks and residue exposure ratings according to an ablation study. The CNN included attention-weight maps which showed how hotspot motifs commonly related to improved stability presented themselves in the model inputs. These findings provide direction for experimentalists who can use them to direct their efforts toward making stability improvements within particular sequence sections thus connecting general artificial intelligence approaches with domain-specific interpretability requirements.

V. LIMITATIONS AND FUTURE WORK

The promising outcomes from this approach face certain performance restrictions. Task-specific stability annotations constrain the training capabilities since many proteins cannot obtain experimental temperature metrics. The VAE generated sequences with high quality yet failed to train models for functional diversity that exceeded thermos ability. Upcoming research will concentrate on generating multi-purpose generative models to attain simultaneous pH and salinity stability along with activity and solubility optimization. Additionally laboratory automation when combined with reinforcement learning and active learning loops would enable performance improvement in a continuous manner.

VI. CONCLUSION

Artificial Intelligence systems have proved their ability to produce efficient thermostable proteins for biotechnological implementation. Our proposed framework uses Random Forests and Convolutional Neural Networks and Graph Neural Networks as well as generative Variational Autoencoder model to create a thorough method which evaluates and enhances and produces protein sequences for increased thermal resistance.

The experimental data confirms the prediction models' accuracy where ensemble methods demonstrate effective correlations against true thermostability measurements. The generative component developed structurally feasible protein variants that also exhibited thermal stability. Experimental testing through laboratory assays demonstrated the practical value of our engineered proteins because they showed elevated melting temperatures but maintained their biological activity.

Traditional protein engineering methods have lower speeds and greater expenses while minimally precise operations but the proposed AI system delivers better results in all three aspects. This innovation demonstrates how AI functions as an evolutionary technological power in designing proteins for synthetic biology and advancing development of enzymes and industrial bio catalytic processes and therapeutic protein applications.

The framework gets its solid foundation from the current study but researchers plan to develop it further through incorporation of additional biophysical properties with functional constraints and reinforcement learning strategies. The updated system will enhance the overall quality and applicability of AI-engineered proteins used in multiple environmental and industrial conditions.

The results demonstrate how artificial intelligence has strengthened its relationship with molecular biology to create new possibilities for quick accurate broad industrial protein engineering solutions in the biotechnology field.

REFERENCES

[1] Bloom, J.D., Labthavikul, S.T., Otey, C.R., Arnold, F.H.: Protein stability promotes evolvability. *Proc. Natl. Acad. Sci. USA* 103, 5869–5874 (2006)

[2] Guerois, R., Nielsen, J.E., Serrano, L.: Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations. *J. Mol. Biol.* 320, 369–387 (2002)

[3] Dehouck, Y., Kwasigroch, J.M., Gilis, D., Rooman, M.: PoPMuSiC 2.1: a web server for the estimation of protein stability changes upon mutation and sequence optimality. *BMC Bioinformatics* 12, 151 (2011)

[4] Capriotti, E., Fariselli, P., Casadio, R.: I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.* 33, W306–W310 (2005)

[5] Aswani, S., Jyotsna, K. A., Pragathi, B., Saidulu, P., Athiraja, A., & Prasannakumar, G. (2024). Detection of Follicular Thyroid Cancer using YOLOv5 Algorithm: A Comparative Analysis with Fuzzy C-Means and Singular Value Decomposition. *International Conference on Advanced Computing and Communication Systems (ICACCS)*.

[6] Sri, M. A., Srikanth, T., Rao, P. S., Kumar, N. S., Athiraja, A., & Rajeswaran, N. (2024). Revolutionizing Tomato Disease Detection: YOLOv5-Based Automated System for Agricultural Productivity and Food Security. *International Conference on Electronics and Sustainable Communication Systems (ICESC)*.

[7] Grace Shalini, T., Susan Shiny, G., Saranya, R., Suresh Babu, P., Kavitha, R., & Atheeswaran, A. (2024). Enhancing Lung Disease Identification with Multimodal Data Fusion and Deep Learning CNN Approach. *International Conference on Smart Electronics and Communication*.

[8] Singh, N., Chozha, R. P., Kumar, R. S., Soujanya, T., Ram, S. S. M., & Athiraja, A. (2025). Adaptive AI Algorithms for Autonomous Crop Harvesting in Variable Terrain Conditions. *International Conference on Information, Implementation, and Innovation in Technology*.

[9] Rakesh, S. K., Punnamchandar, P., N. M., Reddy, R. A., Udaya Kiran, M., & Athiraja, A. (2025). AI-Augmented Data Science Pipelines for Accelerating Knowledge Discovery in Scientific Research. *World Conference on Communication & Computing*.

[10] Yuvaraj, T., Thirumalai, M., Suresh, T. D., Babu, T. S., & Khishe, M. (2025). Dynamic Optimization of Solar DG and Shunt Capacitor Placement to Mitigate the Impact of EV Charging Stations on Power Distribution Network. *Results in Engineering*, 106804.

[11] Abbas, S. H., Ravi, E., Babu, B. M., Pimo, S. J., Kulkarni, P., & Thirumalai, M. (2024). IoTWP: Design and Development of Internet of Things Assisted Weather Prediction Scheme with Advanced Remote Tracking Norms. *International Conference on Power, Energy, Control and Transmission Systems*.

[12] Muthukumaran, D., Patidar, R., Ravi, K. C., Thirumalai, M., Tulasi, R., & Siddiqui, S. T. (2024). Design and Development of LiFi Assisted Intelligent Data Transmission Using Secured Wireless Communication Principles. *International Conference on Power, Energy, Control and Transmission Systems*.

[13] Yuvaraj, T., Devabalaji, K. R., Suresh, T. D., Prabaharan, N., Ueda, S., & Senjyu, T. (2023). Enhancing Indian Practical Distribution System Resilience through Microgrid Formation and Integration of Distributed Energy Resources Considering Battery Electric Vehicle. *IEEE Access*.

[14] Yuvaraj, T., Krishnamoorthy, R., Arun, S., Thanikanti, S. B., & Nwulu, N. (2024). Optimizing Virtual Power Plant Allocation for Enhanced Resilience in Smart Microgrids under Severe Fault Conditions Using the Hunting Prey Optimization Algorithm. *Energy Reports*.

[15] Sriabisha, R., & Yuvaraj, T. (2023). Optimum Placement of Electric Vehicle Charging Station Using Particle Swarm Optimization Algorithm. *International Conference on Electrical Energy Systems*.

[16] Nalinipriya, G., Rama Sree, S., Radhika, K., et al. (2025). Leveraging Explainable Artificial Intelligence for Early Detection and Mitigation of Cyber Threat in Large-Scale Network Environments. *Scientific Reports*.

[17] Ramesh Kumar, R., Nalinipriya, G., Vidyadhari, C., & Elwin, J. G. R. (2024). Deep Joint RP-Net–Based Segmentation Algorithm for Severity Prediction of Brain Tumor. *Journal of Mechanics in Medicine and Biology*.

[18] Nalinipriya, G., Lydia, E. L., Alshenafi, R., Kavuri, R., & Ishak, M. K. (2024). A Two-Tiered Bidirectional Atrous Spatial Pyramid Pooling-Based Semantic Segmentation Model for Landslide Classification Using Remote Sensing Images. *IEEE Access*.

[19] Shrivastav, P., Darji, P., Patel, D., Shah, M., & Shalini, T. G. (2025). Evolutionary Autonomous Car Navigation Using NEAT Algorithm. *International Conference on Sustainable Computing and Data Communication Systems*.

[20] Arsath, R., Shalini, T. G., Yugabharathi, R., & Geetha, P. (2025). Virtual Herbal Garden. *International Conference on Intelligent Communication Technologies and Virtual Mobile Networks*.