Enzyme Classification through Structural Bioinformatics and Advanced Machine Learning Algorithms

Pratham Kaushik[1], Kanwarpartap Singh Gill[1], Nitin Thapliyal[2], Ramesh Singh Rawat[3]

Chitkara University Institute of Engineering and Technology, Chitkara University, Punjab, India[1]

Computer Science & Engineering, Graphic Era Hill University, Dehradun, Uttarakhand, India[2]

Computer Science & Engineering, Graphic Era Deemed to be University, Dehradun, Uttarakhand, India[3]

Corresponding author: Kanwarpartap Singh Gill, Email: kanwarpartap.gill@chitkara.edu.in

This study presents a novel method for enzyme categorization that combines EDA with NN models. Using a dataset of 858,777 annotated amino acid sequences from 10 different species, the model classifies enzymes in a 253,146 sample set, removing those with sequences longer than a certain threshold. X, U, B, and Z are infrequent amino acids that must be omitted during preprocessing in order to make room for B and Z, which are unique to the training set. After 20 epochs, the Neural Network architecture—which includes an embedding layer, bidirectional LSTM layers, and a dense output layer—manages to achieve an encouraging 79% accuracy on the test set. The model's effectiveness across 20 enzyme classes is demonstrated by a comprehensive classification report and confusion matrix. The importance of integrating EDA and NN in bioinformatics and molecular biology is demonstrated by this work, which enhances enzyme categorization approaches. Investigating new features and optimisation techniques to further improve the model is the next step.

Keywords: Enzyme Classification, Amino Acid Sequences, Exploratory Data Analysis, Neural Networks, Bioinformatics.

2024. In Mukesh Saraswat & Rajani Kumari (eds.), *Applied Intelligence and Computing*, 33–40. Computing & Intelligent Systems, SCRS, India. DOI: https://doi.org/10.56155/978-81-955020-9-7-4

Pratham Kaushik[1], Kanwarpartap Singh Gill[1], Nitin Thapliyal[2], Ramesh Singh Rawat[3]

1 Introduction

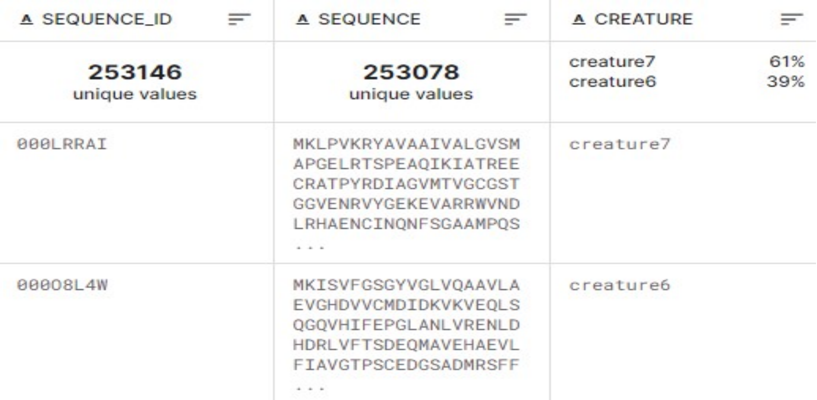
As catalysts that control critical biochemical reactions, enzymes are vital to many biological processes. Medical, bioengineering, and drug discovery advancements depend on our ability to comprehend enzyme structure, function, and evolution. From allostery and catalysis to sustainable manufacturing techniques and genome editing technologies, this introduction covers a wide range of topics that have recently emerged in the field of enzyme studies. A new paradigm for allostery and catalysis in this family of enzymes is introduced by Pan's study on the Rhizophagus irregularis fungus SAMHD1 ortholog [1]. Potentially useful in fields such as agriculture and environmental science, the research elucidates the complex mechanics underpinning enzyme action. Additionally, a scalable and succinct synthesis strategy for a critical intermediate to Belzutifan was presented by Cheung-Lee et al. [2]. They engineered hydroxylase activity, selectivity, and stability. These results demonstrate the promise of enzyme engineering for the pharmaceutical sector and add to our understanding of how to synthesise pharmaceutical intermediates. The significance of eco-friendly procedures in pharmaceutical manufacture is highlighted in the full description of the development of a sustainable and green manufacturing method for Belzutifan by DiRocco et al. [3]. Sustainable procedures are becoming more important in many businesses, and this fits right in with that trend. Exploring the function of TET enzymes in the immunological system, López-Moyado et al. go beyond small molecules to uncover links between DNA demethylation, inflammation, and cancer [4]. Enzymes have several roles, and this study shows that, which has therapeutic implications. The creation of enzyme-specific chemical probes for histone deacetylase substrate profiling using high-throughput technology is the primary focus of Seidel's research [5]. With possible implications for medication development, this work adds to our knowledge of enzyme-mediated epigenetic control. The multidisciplinary character of enzyme research is further demonstrated by Yang et al., who use proteomic and molecular docking studies to seek for patulin-degrading enzymes in Saccharomyces cerevisiae [6]. The groundbreaking potential of enzyme-based technologies in genetic modification is demonstrated by Badon et al.'s discussion of the recent genome editing applications of CRISPR-Cas12 and the OMEGA system [7]. A clinical example of the merging of enzyme-related research with modern computational methodology is Reshan et al.'s use of deep learning techniques for the identification of pneumonia from chest X-ray images [8]. The many functions of enzymes in relation to diet and health are illuminated by Lang et al.'s study of polyphenol categorization and antioxidant tests [9]. Lastly, demonstrating the practical uses of enzyme research in diagnostic technologies, Gill et al. [10] propose an effective VGG19 framework for malaria identification in blood cell pictures. All of these new discoveries put enzyme research in context with its far-reaching and ever-changing effects on other areas of science, paving the way for more investigation and breakthroughs in the future.

2 Literature

Douglas et al. [11] investigate the function of enzyme recognition of amino acids in the genesis of primordial genetic codes and its evolutionary relevance. This paper lays the groundwork for investigating enzyme categorization using amino acid sequences by illuminating chemical interactions and providing an evolutionary perspective. Classification methods that use Artificial Intelligence (AI) have recently become popular in several fields. In their work from 2022, Gill et al. investigate Smart Shoe Classification [12], and in their study from 2022, they employ the VGG19 model to detect brain tumours [15]. The references included here provide light on many AI applications and give useful background for building the enzyme classification model. The GECKO Toolbox is introduced by Chen et al. [16] for the reconstruction and analysis of enzyme-constrained metabolic models, while biocatalysis is investigated by Zhang et al. [13] for the production of active medicinal components. These investigations add to our knowledge of enzyme function by highlighting the importance of enzymes in a wide range of biological and industrial processes. The effects of atrazine on the black soil bacterial population and extracellular enzymes are studied by Gao et al. [14]. Despite the study's environmental bias, it adds to our knowledge of the complex interplay between extraneous variables, microbial populations, and enzyme activity. In their review, Guati et al. [17] discuss how non-enzymatic electrode properties affect glucose sensing. While the main emphasis is on electrodes that do not include enzymes, the study sheds light on biosensors and enzymatic processes in general, which are relevant to the model for classifying enzymes. 'Qingcui' plum fruit treated with 1-MCP during storage undergoes a comprehensive examination of nutritional quality alterations [18]. The integrated analysis technique and molecular control insights help us grasp complex data in biological systems, even if fruit quality is the main emphasis. The variables that affect the meconium microbiota before and throughout pregnancy have been studied by Turunen et al. [19]. Despite the focus on microbiota, the study's examination of variables influencing biological samples sheds light on the difficulties of dealing with various biological data, such as that which is faced in the enzyme classification job. A study conducted by Qian et al. [20] utilised the AU5800 Biochemical Analyzer in conjunction with AI technology to analyse and diagnose hemolytic material. This study shows how AI may be used in biochemical analysis more generally and how it could function in tandem with the enzyme categorization model. Finally, the literature review synthesises results from several research to offer a thorough grasp of important ideas, methods, and problems in the area. The results of these investigations provide the theoretical framework upon which the enzyme categorization model detailed in this study is based.

3 Input Dataset

There are a total of 858,777 labelled amino acid sequences in the training set and 253,146 in the test set that were utilised for this study. A unique identifier (SEQUENCE_ID), an amino acid sequence (SEQUENCE), the organism from which the sample was derived (CREATURE), and the label matching to the set of results (LABEL) are all components of each sample. Importantly, the dataset only contains sequences that fall within a certain range of lengths; any sequences that surpass this limit will not be considered for model optimisation. Finding and dealing with unusual amino acids is an important part of the preprocessing step. Through analysis of the dataset's amino acid distribution, a numerical code dictionary including 20 standard amino acids is revealed. Fig. 1 shows that the model's interpretability is improved and noise is prevented by excluding the amino acids X, U, B, and Z because of how seldom they occur. In addition, the training set is the only one that contains amino acids B and Z, therefore they are treated differently. Taking this into account prior to training the model guarantees accurate recognition and differentiation of these particular amino acids. Particularly when working with varied biological sequences, such careful preprocessing is necessary to build a strong model that can generalise well to new data. The biological heterogeneity in the dataset is amplified by the 10 creatures included, which poses a challenge to the model in terms of learning patterns particular to each organism while yet being able to generalise to others. The Neural Network model for enzyme categorization is built and tested on this varied dataset, which was prepared with great care.



SEQUENCE_ID	SEQUENCE	CREATURE
253146 unique values	253078 unique values	creature7 61% creature6 39%
000LRRAI	MKLPVKRYAVAAIVALGVSM APGELRTSPEAQIKIATREE CRATPYRDIAGVMTVGCGST GGVENRVYGEKEVARRWVND LRHAENCINQNFSGAAMPQS ...	creature7
00008L4W	MKISVFGSGYVGLVQAAVLA EVGHDVVCMDIDKVKVEQLS QGQVHIFEPGLANLVRENLD HDRLVFTSDEQMAVEHAEVL FIAVGTPSCEDGSADMRSFF ...	creature6

Figure 1. Dataset CSV file type utilized for classification purpose

4 Proposed Methodology

Combining Exploratory Data Analysis (EDA) with a Neural Network (NN) architecture, the proposed enzyme classification methodology successfully captures the subtleties of amino acid sequences and organism-specific changes.

4.1 Exploratory Data Analysis (EDA)

A thorough examination of the dataset is carried out in the first phase to learn about the distribution and properties of amino acid sequences. Twenty standard amino acids are identified via amino acid analysis, with an emphasis on removing those with extremely low concentrations to improve the efficiency of the model. To make sure the model can handle differences in the dataset effectively, we pay extra attention to identifying and processing unusual amino acids (X, U, B, and Z). To further simplify the dataset, we further remove sequences that are longer than a certain threshold, as seen in Figure 2.

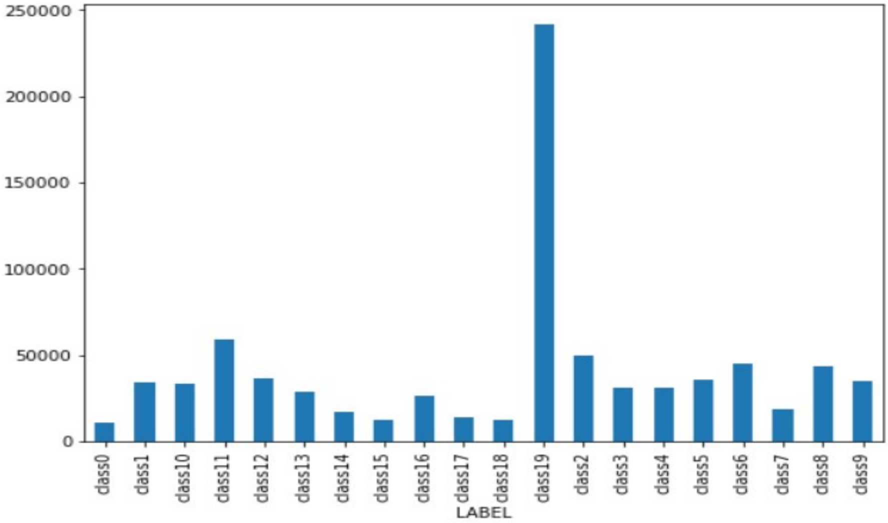
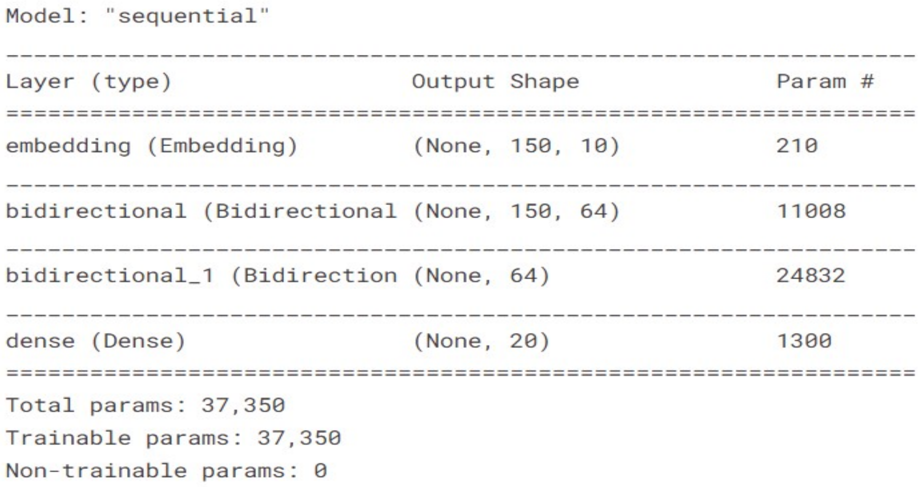


Figure 2. Exploratory Analysis of Data in class wise format

4.2 Neural Network Architecture

Building and training a Neural Network model specifically for enzyme categorization forms the backbone of the suggested technique. A dense representation of amino acids is provided by the model's first embedding layer. Subsequently, two bidirectional Long Short-Term Memory (LSTM) layers efficiently capture the amino acid sequences' sequential dependencies. The last layer is a thick output layer with 20 units, which correspond to the 20 classes of enzymes indicated in Figure 3. During its 20 epochs of training, the model strives for optimal accuracy and generalizability. With 37,350 parameters, the architecture is fine-tuned to efficiently learn the dataset's complicated patterns.



```
Model: "sequential"
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
embedding (Embedding)        (None, 150, 10)           210
_________________________________________________________________
bidirectional (Bidirectional (None, 150, 64)           11008
_________________________________________________________________
bidirectional_1 (Bidirection (None, 64)                24832
_________________________________________________________________
dense (Dense)                (None, 20)                1300
=================================================================
Total params: 37,350
Trainable params: 37,350
Non-trainable params: 0
```

Figure 3. Sequential CNN Model Architecture

4.3 Evaluation Metrics

An in-depth confusion matrix and classification report are part of the performance evaluation. These documents include metrics for each enzyme class, including F1-score, recall, and accuracy. The suggested technique may be better understood and validated with the help of this thorough study, which sheds light on the model's class discrimination capabilities. Incorporating EDA and NN models into enzyme categorization provides a comprehensive strategy that takes into account data exploration and predictive modelling, strengthening the suggested technique as a whole.

5 Results

This research paper's findings section showcases the usefulness of the suggested model in enzyme categorization, as it attained a respectable accuracy of 79% on the test set. The 20 enzyme classes are thoroughly evaluated using the precision, recall, and F1-score metrics, which are presented in the classification report. The accuracy of the model's class predictions and its room for improvement are both shown by the confusion matrix. The model does an excellent job of ignoring uncommon amino acids and organism-specific changes. These outcomes prove that the EDA-Neural Network hybrid method can successfully classify enzymes using amino acid sequences.

5.1 Training and Validation Curve Analysis

The study paper's training curve analysis shows how the model learned more and more over the course of the 20 training epochs. Over the course of the model's lifetime, the accuracy and loss values on the training and validation sets gradually improve while the loss values steadily decline. The model's generalizability is demonstrated when these curves converge, indicating that it successfully fits the training data while also making accurate predictions on unknown data. Figure 4 shows the model's learning dynamics and convergence, and the patterns in the training curves confirm that the neural network design is successful for enzyme categorization.

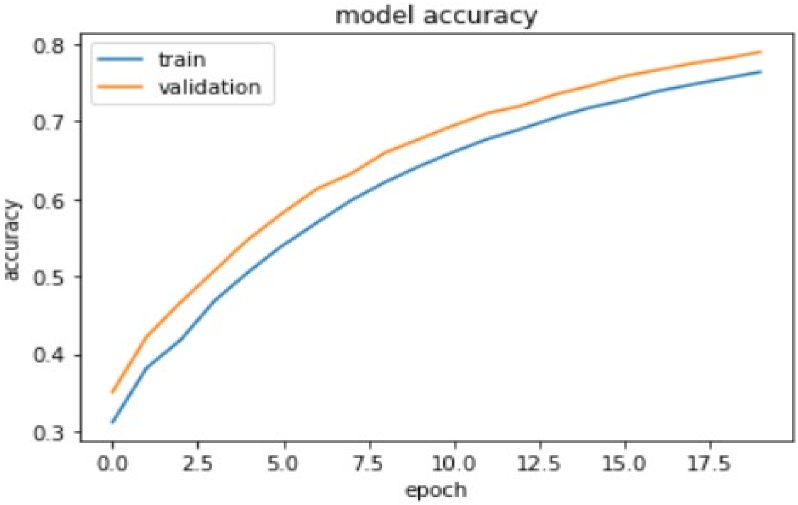
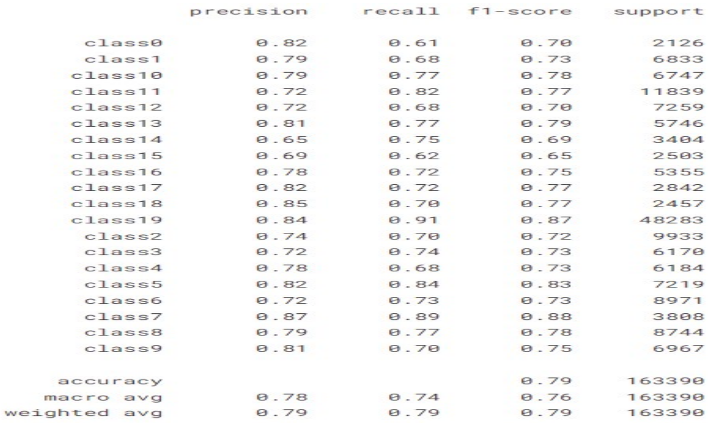


Figure 4. Model Accuracy depiction through graph

5.2 Classification Report Analysis

The model's performance across 20 enzyme classes is thoroughly evaluated in the classification report analysis. The accuracy of optimistic forecasts is reflected in precision, which stands at 79% overall. Sensitivity varies among classes, as seen by the recall score, which measures the capacity to catch true positives and spans from 61% to 91%. Figure 5 shows that the model achieves a well-balanced performance with an average F1-score of 76%, which combines precision and recall. The suggested Enzyme Classification methodology is generally effective, as evidenced by the weighted average accuracy of 79%. Taken as a whole, these metrics highlight the model's predictive power for enzyme classes, which is great for bioinformatics.



```
              precision    recall  f1-score   support

      class0       0.82      0.61      0.70      2126
      class1       0.79      0.68      0.73      6833
     class10       0.79      0.77      0.78      6747
     class11       0.72      0.82      0.77     11839
     class12       0.72      0.68      0.70      7259
     class13       0.81      0.77      0.79      5746
     class14       0.65      0.75      0.69      3404
     class15       0.69      0.62      0.65      2503
     class16       0.78      0.72      0.75      5355
     class17       0.82      0.72      0.77      2842
     class18       0.85      0.70      0.77      2457
     class19       0.84      0.91      0.87     48283
      class2       0.74      0.70      0.72      9933
      class3       0.72      0.74      0.73      6170
      class4       0.78      0.68      0.73      6184
      class5       0.82      0.84      0.83      7219
      class6       0.72      0.73      0.73      8971
      class7       0.87      0.89      0.88      3808
      class8       0.79      0.77      0.78      8744
      class9       0.81      0.70      0.75      6967

    accuracy                           0.79    163390
   macro avg       0.78      0.74      0.76    163390
weighted avg       0.79      0.79      0.79    163390
```

Figure 5. Classification Report Analysis

5.3 Confusion Matrix Analysis

A comprehensive review of the model's performance across all 20 enzyme classes is given by the confusion matrix. There is a row for the actual class and a column for the anticipated class. As a demonstration of the model's accuracy in instance classification, the diagonal components depict the real positive predictions for each class. Elements that do not fit neatly on the graph are misclassifications, which provide information about the model's weaknesses. As seen in Figure 6, improving the model to make more accurate predictions across various enzyme classes is possible by analysing the confusion matrix, which helps to understand the model's strengths and shortcomings.

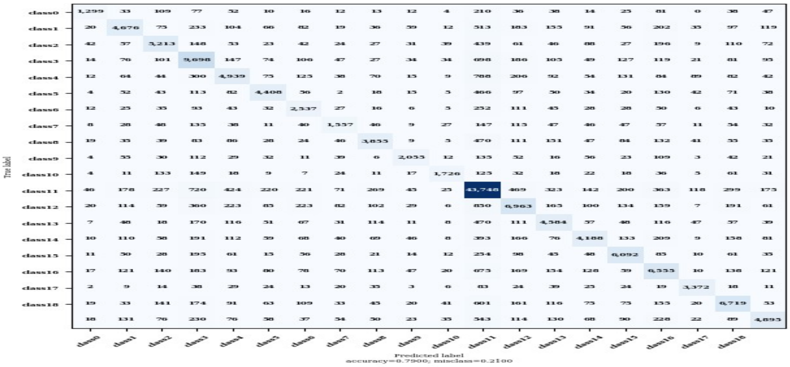


Figure 6. Confusion Matrix Analysis

6 Conclusion

This study concludes with an all-encompassing method for enzyme categorization that uses a Neural Network (NN) model in conjunction with Exploratory Data Analysis (EDA). With a remarkable 79% accuracy rate, the model shows promise as an enzyme class predictor, having been trained on a large dataset of amino acid sequences. The model's robustness is enhanced by factors such as the elimination of uncommon amino acids and the consideration of changes peculiar to each organism. The model's competence in differentiating between enzyme classes is demonstrated by the classification report and confusion matrix, which exhibit respectable metrics for recall, accuracy, and F1-score. To get a more complex knowledge of enzyme categorization, future studies might investigate the effect of sequence length, investigate different deep learning architectures, and add more biological characteristics to the current model, which works adequately. In sum, the offered technique lays the groundwork for future advancements in bioinformatics enzyme categorization, opening the door to better biological insights and more complex models.

References

[1] Pan, L., The SAMHD1 Ortholog from the Fungus Rhizophagus irregularis Presents a New Paradigm for Allostery and Catalysis in This Enzyme Class (Doctoral dissertation, Brandeis University, Graduate School of Arts & Sciences).

[2] Cheung-Lee, W.L., Kolev, J.N., McIntosh, J.A., Gil, A.A., Pan, W., Xiao, L., Velásquez, J.E., Gangam, R., Winston, M.S., Li, S. and Abe, K., 2024. Engineering Hydroxylase Activity, Selectivity, and Stability for a Scalable Concise Synthesis of a Key Intermediate to Belzutifan. Angewandte Chemie International Edition, p.e202316133.

[3] DiRocco, D.A., Zhong, Y.L., Le, D.N., McCann, S.D., Hethcox, J.C., Kim, J., Kolev, J.N., Kosjek, B., Dalby, S.M., McMullen, J.P. and Gangam, R., 2024. Evolution of a Green and Sustainable Manufacturing Process for Belzutifan: Part 1— Process History and Development Strategy. Organic Process Research & Development.

[4] López-Moyado, I.F., Ko, M., Hogan, P.G. and Rao, A., 2024. TET Enzymes in the Immune System: From DNA Demethylation to Immunotherapy, Inflammation, and Cancer. Annual Review of Immunology, 42.

[5] Seidel, J., 2024. Development of enzyme-specific chemical probes for high-throughput substrate profiling of histone deacetylases (Doctoral dissertation, Universität Tübingen).

[6] Yang, C., Zhang, Z. and Peng, B., 2024. New insights into searching patulin degrading enzymes in Saccharomyces cerevisiae through proteomic and molecular docking analysis. Journal of Hazardous Materials, 463, p.132806.

[7] Badon, I.W., Oh, Y., Kim, H.J. and Lee, S.H., 2024. Recent application of CRISPR-Cas12 and OMEGA system for genome editing. Molecular Therapy, 32(1), pp.32-43.

[8] Reshan, M.S.A., Gill, K.S., Anand, V., Gupta, S., Alshahrani, H., Sulaiman, A. and Shaikh, A., 2023, May. Detection of Pneumonia from Chest X-ray Images Utilizing MobileNet Model. In Healthcare (Vol. 11, No. 11, p. 1561). MDPI.

[9] Lang, Y., Gao, N., Zang, Z., Meng, X., Lin, Y., Yang, S., Yang, Y., Jin, Z. and Li, B., 2024. Classification and antioxidant assays of polyphenols: A review. Journal of Future Foods, 4(3), pp.193-204.

[10] Gill, K.S., Anand, V. and Gupta, R., 2023, August. An Efficient VGG19 Framework for Malaria Detection in Blood Cell Images. In 2023 3rd Asian Conference on Innovation in Technology (ASIANCON) (pp. 1-4). IEEE.

[11] Douglas, J., Bouckaert, R., Carter Jr, C.W. and Wills, P.R., 2024. Enzymic recognition of amino acids drove the evolution of primordial genetic codes. Nucleic Acids Research, 52(2), pp.558-571.

[12] Gill, K.S., Sharma, A., Anand, V. and Gupta, R., 2023, May. Smart Shoe Classification Using Artificial Intelligence on EfficientnetB3 Model. In 2023 International Conference on Advancement in Computation & Computer Technologies (InCACCT) (pp. 254-258). IEEE.

[13] Zhang, N., Domínguez de María, P. and Kara, S., 2024. Biocatalysis for the Synthesis of Active Pharmaceutical Ingredients in Deep Eutectic Solvents: State-of-the-Art and Prospects. Catalysts, 14(1), p.84.

[14] Gao, T., Tian, H., Xiang, L., Wang, Z., Fu, Y., Shi, J., Wen, X., Jiang, X., He, W., Hashsham, S.A. and Wang, F., 2024. Characteristics of bacterial community and extracellular enzymes in response to atrazine application in black soil. Environmental Pollution, 343, p.123286.

[15] Gill, K.S., Sharma, A., Anand, V. and Gupta, R., 2022, December. Brain Tumor Detection using VGG19 model on Adadelta and SGD Optimizer. In 2022 6th International Conference on Electronics, Communication and Aerospace Technology (pp. 1407-1412). IEEE.

[16] Chen, Y., Gustafsson, J., Tafur Rangel, A., Anton, M., Domenzain, I., Kittikunapong, C., Li, F., Yuan, L., Nielsen, J. and Kerkhoven, E.J., 2024. Reconstruction, simulation and analysis of enzyme-constrained metabolic models using GECKO Toolbox 3.0. Nature Protocols, pp.1-39.

[17] Guati, C., Gomez-Coma, L., Fallanza, M. and Ortiz, I., 2024. Progress on the influence of non-enzymatic electrodes characteristics on the response to glucose detection: A review (2016–2022). Reviews in Chemical Engineering, 40(1), pp.123-148.

[18] Du, L., Kou, L., Liu, D., Hu, W., Yu, Y., Luo, G., Lai, B. and Cai, J., 2024. Integrated analysis of nutritional quality changes and molecular regulation in 'Qingcui'plum fruit treated with 1-MCP during storage. Postharvest Biology and Technology, 207, p.112591.

[19] Turunen, J., Tejesvi, M.V., Paalanne, N., Pokka, T., Amatya, S.B., Mishra, S., Kaisanlahti, A., Reunanen, J. and Tapiainen, T., 2024. Investigating prenatal and perinatal factors on meconium microbiota: a systematic review and cohort study. Pediatric Research, 95(1), pp.135-145.

[20] Qian, J., Song, T., Zhang, Q., Cai, G. and Cai, M., 2024. Analysis and Diagnosis of Hemolytic Specimens by AU5800 Biochemical Analyzer Combined With AI Technology. Frontiers in Computing and Intelligent Systems, 6(3), pp.100-3.