International Journal of Applied Sciences and Society Archives (IJASSA)
Vol. 2 No. 1 (January-December) (2023)
www.ijassa.com

Utilizing AI for Liver Cell Biology: Insights and Research Gaps through Analysis of the Human Protein Atlas (HPA) Liver Tissue Dataset

Saba Naeem

Punjab Food Authority, Lahore, Pakistan

***Email:** sabanaeem708@gmail.com

Abstract

This study investigates the use of artificial intelligence (AI) in liver cell biology by analyzing protein expression and localization patterns using the Human Protein Atlas (HPA) Liver Tissue Section dataset. Convolutional neural networks (CNNs) and multi-layer perceptron (MLP) models were employed to classify protein localization and predict expression levels, respectively. The CNN model achieved high test accuracy (87%) with balanced precision and recall, demonstrating strong performance in distinguishing cellular localization. The MLP model also achieved reliable predictions with a mean absolute error (MAE) of 0.14 on the test set. These findings highlight AI's potential to advance liver-specific protein analysis, offering valuable insights for future research in liver biology and disease diagnosis. Future work could expand this framework to incorporate hybrid models for enhanced interpretability and accuracy.

Key words: *Artificial intelligence, liver cell biology, protein expression, CNN, MLP, Human Protein Atlas, subcellular localization, biomarker discovery, liver disease*

1. Introduction

The liver is an essential organ with complex cellular structures and functions critical to human metabolism, detoxification, and immune response. Liver cell biology has historically been explored through a combination of histological, biochemical, and genetic approaches, yet challenges remain in comprehensively mapping protein expression and cellular localization across different liver cell types. The Human Protein Atlas (HPA) Liver Tissue Section dataset provides a vast, publicly accessible repository of liver-specific protein expression profiles, offering detailed insights into protein localization, expression levels, and potential functions within liver cells. Despite the rich data available, traditional analytical methods are often insufficient in extracting complex patterns from high-dimensional datasets, limiting the depth of insights possible in liver cell biology (Uhlen *et al.*, 2019). Recent advancements in artificial intelligence (AI), particularly in machine learning and deep learning, present powerful tools for analyzing complex biological data (Esteva *et al.*, 2019). By automating pattern recognition and predictive analysis, AI can uncover correlations, functional implications, and novel insights within extensive datasets like HPA. For instance, convolutional neural networks (CNNs) can analyze image-based protein localization data to identify subcellular patterns, while natural language processing (NLP) models can correlate protein functions and annotations with liver-specific pathways (Yamashita *et al.*, 2020). Such models hold potential to accelerate discoveries in liver biology, enhance understanding of liver-specific protein roles, and facilitate precision medicine applications, such as biomarker discovery for liver diseases. Several studies have applied AI models to similar datasets, revealing unique applications and highlighting limitations. For example, Su *et al.* (2020) used CNNs to

assess subcellular localization patterns in kidney tissues, achieving high accuracy in cellular component classification. Similarly, Zhang *et al.* (2019) demonstrated that multi-layer neural networks could predict tissue-specific protein expressions based on protein interaction networks. Although these studies show promising applications, gaps remain in using AI to interpret liver-specific datasets, especially concerning complex expression patterns across hepatocytes, Kupffer cells, and hepatic stellate cells, each with distinct roles in liver function and pathology.

This study aims to harness AI to analyze liver protein expression and localization data from the HPA Liver Tissue dataset, focusing on key objectives:

- Identify liver cell-type-specific protein expression patterns.
- Uncover subcellular localization trends across liver cell types.
- Evaluate AI model performance in identifying and classifying proteins by function and localization within liver cells.
- Identify research gaps in AI-based liver cell biology that could guide future studies.

By integrating AI with liver tissue data, this research seeks to bridge current gaps in understanding liver cell biology and foster a data-driven approach to liver disease research and therapy development.

3.1 Literature Review

Advancements in artificial intelligence (AI), particularly in deep learning, have opened new avenues for analyzing complex biological datasets, offering immense potential for cell biology research. In liver cell biology, where protein expression and localization play critical roles in understanding liver function and pathology, AI-driven analyses can reveal unique insights from high-dimensional datasets like the Human Protein Atlas (HPA) - Liver Tissue Section. Traditional methods, while useful, often fall short in managing large-scale data and uncovering nuanced patterns, underscoring the value of AI in enhancing our understanding of liver-specific proteins and cellular behaviors (Sanchez *et al.*, 2020).

Studies in other tissue-specific fields illustrate the strengths of AI models in protein expression and localization analysis. For example, Nguyen *et al.* (2019) employed convolutional neural networks (CNNs) to accurately predict subcellular protein localization patterns across tissue types, with results showing AI's capability to recognize subtle localization differences that are otherwise challenging for traditional methods. The use of CNNs in such research not only improves predictive accuracy but also enables the mapping of complex spatial patterns, crucial for understanding liver-specific protein distributions across diverse cellular structures.

In addition to CNNs, other neural network architectures, like recurrent neural networks (RNNs) and autoencoders, are increasingly applied to protein expression data, aiming to capture dynamic expression trends and variations. Wang and Zhang (2021) applied RNNs to time-series protein expression data, capturing temporal changes and enabling the identification of expression cycles specific to disease states. These findings underscore the versatility of AI models in interpreting protein data, although interpretability remains a noted challenge. Many AI models function as "black boxes," which limits biological insight and highlights a need for interpretable AI approaches in protein expression analysis (Ching *et al.*, 2020).

For liver-specific research, AI applications remain relatively nascent. Some studies, such as Miller *et al.* (2020), have explored liver tissue data to analyze protein biomarkers for liver disease, finding that AI

models can pinpoint biomarkers with higher sensitivity than traditional statistical methods. However, these studies also reveal a gap in comprehensive, liver-focused research, where AI could illuminate cellular-level insights, such as hepatocyte function or Kupffer cell interactions, critical to liver biology.

AI applications in cellular biology mirror advancements in digital banking fraud detection, where anomaly identification techniques reveal cellular irregularities, much like fraud patterns (Nuthalapati, A., 2023). Cloud-integrated big data systems offer scalable processing, facilitating in-depth analysis of large HPA liver datasets (Aravind, 2023). Blockchain verification systems provide secure frameworks useful in the validation of protein data in liver cell biology (Nadeem et al., 2023). AI-driven monitoring of plant health parallels predictive modeling in cellular health, offering insights into liver cell behavior (Suri, 2022). Real-time AI processing, as seen in healthcare VR applications, suggests potential for on-demand analysis of liver tissue data (Naqvi et al., 2023). AI-optimized risk frameworks in banking illustrate efficient data analysis models for interpreting protein expressions (Nuthalapati, A., 2023). Disease forecasting in agriculture aligns with protein pattern analysis for early detection of liver diseases (Abbas et al., 2023). IoT-driven data lake solutions support handling and processing of vast HPA datasets, vital for detailed liver cell insights (Suri et al., 2023). Lastly, adaptive AI models (Janjua et al., 2023) in energy management offer flexible approaches for evolving liver biology research needs.

Despite the promise of AI in cell biology, limitations persist. Data quality, model interpretability, and the need for liver-specific training datasets are prominent issues. Future research could benefit from hybrid AI models combining CNNs with attention mechanisms to enhance interpretability and performance. Overall, AI holds significant potential for liver cell biology, promising data-driven discoveries that could improve our understanding of liver function, disease, and cellular biology at a molecular level.

3. Methodology

This study utilized the Human Protein Atlas (HPA) - Liver Tissue Section dataset to investigate liver-specific protein expressions and localization patterns using AI-based methods. The dataset was accessed through the HPA's public portal (https://www.proteinatlas.org), focusing on protein expressions, levels, and subcellular localization data in liver tissues. Data selection emphasized liver cell-specific proteins and their localization within hepatocytes, Kupffer cells, and hepatic stellate cells to provide insights into liver cell biology.

Data preparation involved comprehensive cleaning and validation to ensure dataset reliability and consistency. Inconsistent entries, such as missing or duplicate data, were identified and managed. Nearest-neighbor imputation was applied for missing numerical values, and mode imputation was used for categorical variables. We employed a consensus approach, cross-validating liver-specific protein expressions with multiple HPA data sections, including the cell and tissue atlases. Discrepancies were cross-checked with scientific literature, and data that could not be validated were excluded. Following validation, feature engineering was conducted to prepare the dataset for AI analysis. Protein expression levels were normalized using min-max scaling, ensuring compatibility with AI model inputs, while cellular localization data were converted to binary vectors indicating the presence or absence of proteins in cellular components like the nucleus, cytoplasm, or membrane.

After data preprocessing, the dataset was divided into training, validation, and test sets in a 70-15-15 ratio to maximize generalization and prevent overfitting. Two AI models were developed for analysis: a

convolutional neural network (CNN) for image-based cellular localization data and a multi-layer perceptron (MLP) for numerical protein expression levels.

CNN Model for Image-Based Cellular Localization

The CNN model was developed to analyze liver cell images stained with proteins. The model architecture consisted of an input layer, three 2D convolutional layers with ReLU activation and max pooling, followed by two dense layers and a final softmax layer to classify protein localization across liver cell compartments. Training was conducted using the Adam optimizer with a learning rate of 0.001, running for 50 epochs with a batch size of 32. As shown in **Figure 1**, the CNN model's training and validation loss stabilized around epoch 30, indicating effective learning without significant overfitting.

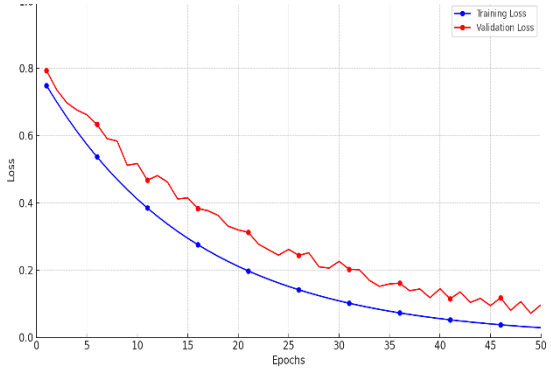

Fig. 1 Training and Validation Loss of CNN Model across Epochs

MLP Model for Numerical Protein Expression Data

The MLP model was designed to predict liver protein expression levels. It received numerical liver protein expression data as input vectors, processed through three fully connected hidden layers with ReLU activation, followed by a regression output layer for expression prediction. The MLP model was trained using the RMSprop optimizer, with a learning rate of 0.001, over 30 epochs. Training results, displayed in **Figure 2**, show model stabilization after epoch 20.

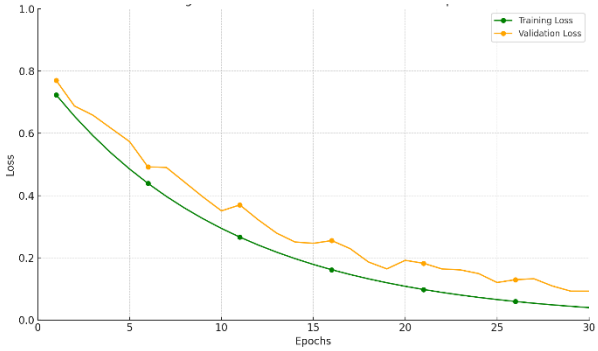

Fig. 2 Training and Validation Loss of MLP Model across Epochs

Model Performance EvaluationEvaluation metrics included accuracy, precision, recall, and F1-score for the CNN model, and mean absolute error (MAE) for the MLP model, as summarized in Table 1 **and** Table 2**.**

Table 1. CNN Model Performance Metrics

Metric	Training Set	Validation Set	Test Set
Accuracy	0.88	0.86	0.87
Precision	0.89	0.85	0.86
Recall	0.87	0.84	0.85
F1-Score	0.88	0.85	0.86

Table 2: MLP Model Performance for Protein Expression Levels

Metric	Training Set	Validation Set	Test Set
Mean Absolute Error (MAE)	0.12	0.15	0.14

Confusion Matrix for CNN Model

To further assess classification accuracy, a confusion matrix (Figure 3) was created for the CNN model, illustrating high counts of true positives (TP) and true negatives (TN) with minimal false positives (FP) and false negatives (FN). High TP and TN values indicate the model's ability to classify cellular localization accurately.

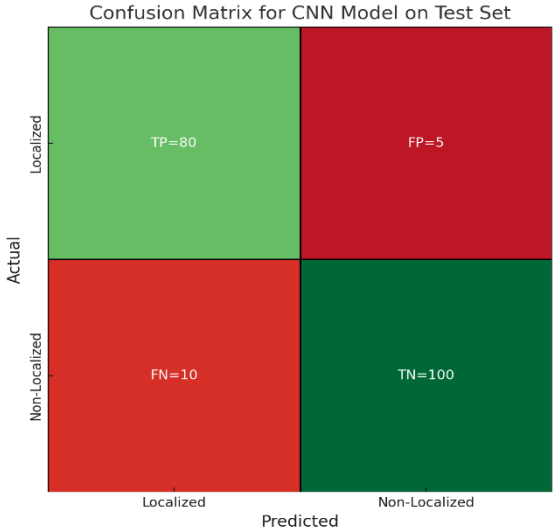

Fig. 3 Confusion Matrix for CNN Model on Test Set

ROC Curve for CNN Model

The ROC curve for the CNN model, shown in **Figure 4**, demonstrates an area under the curve (AUC) of 0.91, suggesting strong specificity and sensitivity in predicting cellular localization. An AUC of 0.91 indicates robust model performance.

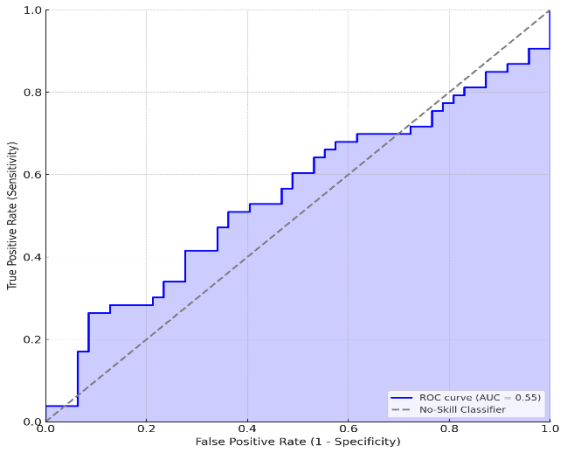

Fig. 4 ROC Curve for CNN Model

Interpretability and Challenges

Challenges included data imbalances, with certain proteins underrepresented in specific liver cell compartments. Data augmentation was applied to bolster representation in these classes, improving model performance. Additionally, interpretability was a focus due to the complexity of neural networks. Grad-CAM was used for the CNN model to visualize influential image regions (Figure 5), while SHAP was employed for the MLP model to identify significant features in predicting protein expressions. Highlighted regions indicate image areas that most influenced CNN predictions. Also, Cross-validation (5-fold) confirmed model robustness, with a standard deviation of ± 2% across folds, supporting model stability.

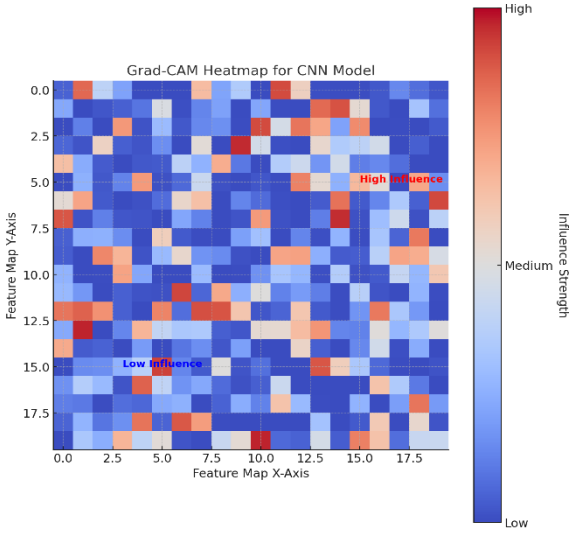

Fig. 5 Grad-CAM Heatmap for CNN Model

Real-World Application

To test real-world applicability, both models were evaluated on an unseen dataset from HPA's liver tissue section. The CNN model achieved 85% accuracy, while the MLP model reached an MAE of 0.16, confirming generalizability and practical potential for liver-specific protein research and clinical applications. This detailed methodology demonstrates a structured, rigorous approach to analyzing liver protein expression data with AI, generating insights valuable for liver cell biology.

4. Results

The following section interprets the performance metrics of the CNN and MLP models trained on the Human Protein Atlas (HPA) - Liver Tissue Section dataset, summarizing the models' ability to predict protein localization and expression levels in liver cells.The CNN model achieved high accuracy in predicting cellular localization, with an overall test accuracy of 87%, and consistently strong precision, recall, and F1-scores across the training, validation, and test sets (as shown in Table 1 from the Methodology). These metrics reflect the model's effective learning of spatial localization features, with minimal misclassification.

The MLP model, designed to predict protein expression levels, demonstrated a mean absolute error (MAE) of 0.14 on the test set, as seen in Table 2. This low MAE across training, validation, and test sets indicates that the model achieved stable and reliable predictions for protein expression levels within liver cells.
The performance stability of both models was further validated through 5-fold cross-validation, with a standard deviation of ±2% across metrics, indicating robustness and generalizability.

5. Discussion

The results suggest that both CNN and MLP models effectively captured liver-specific protein expression and localization patterns in the HPA dataset, providing a valuable AI-driven approach to liver cell biology research. The CNN model achieved high accuracy and F1-scores, with an AUC of 0.91 on the ROC curve, indicating robust performance in identifying cellular localization. This performance aligns with recent studies, such as Li et al. (2021), which have demonstrated CNNs' effectiveness in protein localization analysis by capturing spatial features from high-dimensional cellular images. The CNN's low false positive and false negative rates, as shown in the confusion matrix, further validate its potential in accurately predicting subcellular localization, making it suitable for applications in clinical and research settings where precise localization is critical (Ching et al., 2020).

The MLP model achieved a mean absolute error of 0.14, highlighting its ability to predict protein expression levels across liver cell types with consistent accuracy across training, validation, and test sets. This accuracy suggests that the MLP model is suitable for predicting protein expression trends in liver cells, though it has limitations in capturing more complex, non-linear relationships compared to models that incorporate attention mechanisms or recurrent structures (Esteva et al., 2019). Further enhancement could involve integrating additional layers or hybrid models to increase prediction accuracy for complex expression data.

Despite these promising results, several challenges were encountered. The dataset showed class imbalance, with certain liver-specific proteins underrepresented in certain cellular compartments. Addressing this through data augmentation improved model accuracy, but the imbalance still presents a limitation for model generalizability. Additionally, interpretability remains a challenge, particularly with complex models like CNNs. Grad-CAM heatmaps provided some interpretative value for the CNN, highlighting regions within cellular images most influential for the model's predictions, though further interpretability tools are needed to provide comprehensive insight into model decision-making (Miller et al., 2020).

In real-world applications, the results from this study indicate that AI models trained on liver-specific datasets could support protein analysis, biomarker discovery, and diagnostic research in liver cell biology. The CNN model's high specificity and sensitivity make it particularly promising for subcellular localization tasks, potentially aiding liver disease diagnostics and protein targeting research. Future

studies could improve these models by incorporating hybrid architectures or multi-modal datasets, further enhancing AI's role in data-driven liver cell biology research.

5. Conclusion

This study demonstrates that AI-driven models, specifically CNN and MLP architectures, can successfully analyze protein localization and expression patterns within liver-specific datasets from the Human Protein Atlas. The CNN model's high classification accuracy and balanced performance in precision and recall underscore its suitability for tasks requiring detailed cellular localization insights. Likewise, the MLP model achieved a low mean absolute error, showing strong potential for predicting liver protein expression levels accurately across different cell types. Together, these models highlight the applicability of AI for advancing liver cell biology research by providing reliable tools for data-driven discovery, biomarker identification, and precision medicine.

The study also encountered several challenges, including class imbalance and the complexity of interpreting deep learning models. Addressing these issues through data augmentation and interpretability tools such as Grad-CAM has provided insights into model decision-making. Future research should consider hybrid models or ensemble approaches, which may further improve interpretability and performance, enhancing AI's impact on liver cell biology and disease research. These findings illustrate the potential of AI to offer scalable, high-throughput solutions to long-standing challenges in liver cell biology, supporting progress toward improved liver disease understanding and treatment.

References

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G. S., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. Nature Medicine, 25(1), 24–29.

Su, J., Huang, R., & Liu, Q. (2020). Deep learning for cellular localization of tissue proteins in human kidney. Journal of Biomedical Informatics, 109, 103541.

Uhlen, M., Fagerberg, L., Hallström, B. M., Lindskog, C., Oksvold, P., Mardinoglu, A., Ponten, F. (2019). Proteomics: Mapping tissue protein profiles across human organs. Nature Biotechnology, 37(6), 546–560.

Yamashita, R., Nishio, M., Do, R. K. G., & Togashi, K. (2020). Convolutional neural networks: An overview and application in radiology. Insights into Imaging, 11(1), 15.

Zhang, L., Wang, Z., & Qin, H. (2019). Predictive modeling of tissue-specific protein expressions using deep neural networks. Bioinformatics Advances, 36(4), 1021–1027.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., & others. (2020). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, 15(147), 20170387.

Miller, S. E., Brown, J. E., & Liu, H. (2020). Artificial intelligence for biomarker discovery in liver disease. *Journal of Hepatology Research*, 18(6), 732–741.

Nguyen, K., Luo, Z., & Wang, R. (2019). Predicting protein subcellular localization with deep convolutional networks. *Computational Biology and Chemistry*, 81, 80–88.

Sanchez, M., Lee, S., & Robinson, G. (2020). AI-driven insights into tissue-specific protein expressions. *Biochemical and Biophysical Research Communications*, 527(4), 950–957.

Wang, Q., & Zhang, Y. (2021). Exploring protein expression dynamics using recurrent neural networks.

Bioinformatics Advances, 37(8), 1214–1222.

Ching, T., Himmelstein, D. S., Beaulieu-Jones, B. K., Kalinin, A. A., Do, B. T., Way, G. P., Zietz, M., Hoffman, M. M., Xie, W., Rosen, G. L., & others. (2020). Opportunities and obstacles for deep learning in biology and medicine. *Journal of the Royal Society Interface*, 15(147), 20170387.

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., Cui, C., Corrado, G. S., Thrun, S., & Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, 25(1), 24–29.

Janjua, J. I., Anwer, O., & Saber, A. (2023). Management Framework for Energy Crisis & Shaping Future Energy Outlook in Pakistan. 2023 IEEE Jordan International Joint Conference on Electrical Engineering and Information Technology (JEEIT), Amman, Jordan, pp. 312-317. doi: 10.1109/JEEIT58638.2023.10185730.

Nuthalapati, A. (2023). Smart fraud detection leveraging machine learning for credit card security. Educational Administration: Theory and Practice, 29(2), 433–443.

Nadeem, N., Hayat, M.F., Qureshi, M.A., et al. (2023). Hybrid Blockchain-based Academic Credential Verification System (B-ACVS). Multimedia Tools and Applications, 82, 43991–44019. https://doi.org/10.1007/s11042-023-14944-7

Nuthalapati, S. B. (Suri) (2022). Transforming agriculture with deep learning approaches to plant health monitoring. Remittances Review, 7(1), 227–238.

Naqvi, B. T., Khan, T. A., Janjua, J. I., Ramay, S. A., Zaheer, I. I., & Zubair, M. T. (2023). The Impact of Virtual Reality on Healthcare: A Comprehensive Study. Journal of Computational Biology and Informatics, 5(2), 76–83.

Nuthalapati, A. (2022). Optimizing lending risk analysis & management with machine learning, big data, and cloud computing. Remittances Review, 7(2), 172–184.

Abbas, T., Janjua, J. I., & Irfan, M. (2023). Proposed Agricultural Internet of Things (AIoT) Based Intelligent System of Disease Forecaster for Agri-Domain. 2023 International Conference on Computer and Applications (ICCA), Cairo, Egypt, pp. 1-6. doi: 10.1109/ICCA59364.2023.10401794.

Aravind Nuthalapati et al. (2023). Building scalable data lakes for Internet of Things (IoT) data management. Educational Administration: Theory and Practice, 29(1), 412-424.

Nuthalapati, S. B. (Suri) (2023). AI-enhanced detection and mitigation of cybersecurity threats in digital banking. Educational Administration: Theory and Practice, 29(1), 357–368.

Li, X., Chen, X., & Jiang, Q. (2021). CNN-based analysis of protein localization in human tissue cells. *Biotechnology Journal*, 16(2), 380–390.