International Journal of Computing and Digital Systems
ISSN (2210-142X)
Int. J. Com. Dig. Sys. , No. (Mon-20..)

http://dx.doi.org/10.12785/ijcds/XXXXXX

Liver Disease Diagnostics with Explainable AI and Deep Learning

Maheen Islam[1], Nishat Vasker[1] and Mahamudul Hasan[1]

[1]*Department of Computer Science and Engineering, East West University, Dhaka, Bangladesh*

Received Mon. 20, Revised Mon. 20, Accepted Mon. 20, Published Mon. 20

Abstract: The transformative potential of Artificial Intelligence (AI) in medical diagnostics is hampered by the ”black-box” challenge, where the complex workings of deep learning models obscure the clarity necessary for clinical trust. This research confronts the opacity of AI systems by integrating Explainable Artificial Intelligence (XAI) in liver disease diagnosis, aiming to enhance interpretability and foster healthcare professionals' confidence in AI-driven decisions. This study focuses on whether XAI can demystify the predictive mechanics of deep learning models in medical imaging and examines its effect on the trust and reliability perceived by healthcare professionals. Employing empirical methodologies, a deep learning model was developed for diagnosing liver diseases from medical imaging data, featuring XAI for transparency. The implementation yielded a deep learning model with an 81% accuracy rate, achieving considerable interpretability through SHAP (SHapley Additive exPlanations) values without compromising diagnostic performance. The integration of XAI provided insights, with features like Alkaline Phosphatase showing a significant mean SHAP value of +0.07, underscoring its predictive prominence. The inclusion of XAI in AI diagnostics not only clarifies the decision-making process but also enhances user trust, potentially leading to broader clinical application. The originality of this work lies in its approach to fusing deep learning with XAI, contributing to the progressive vision of transparent, personalized medicine. This research can aid practitioners in leveraging AI for liver disease diagnosis, advancing the domain of biomedical AI.

Keywords: Explainable AI, Explainable AI in liver disease, xai with deep learning, Explainable AI-based medical diagnostic,XAI classification, SHAP analysis, SHAP analysis in liver disease

1. INTRODUCTION

The advent of Artificial Intelligence (AI) in the medical domain has revolutionized the landscape of disease diagnosis and prognosis. With the capability to analyze complex biomedical data, AI algorithms, particularly those based on deep learning, have demonstrated remarkable success in identifying and predicting various health conditions. Despite these advancements, the opaqueness of AI decision-making processes poses significant challenges, often termed as the ”black-box” phenomenon, which impedes their trustworthiness and clinical acceptance. The motivation of this research stems from the necessity to bridge this gap by leveraging Explainable Artificial Intelligence (XAI) to provide transparency and interpretability in AI-driven medical diagnostics.

This study is driven by pivotal questions that address the core concerns in the integration of AI within medical diagnostics: How can XAI contribute to enhancing the interpretability of deep learning models in medical imaging?What is the impact of XAI on the trust and reliability perceived by healthcare professionals towards AI-driven diagnostic systems?

The primary objectives of this research are To develop a deep learning model for the diagnosis of liver diseases using medical imaging.To implement XAI methodologies that facilitate the understanding of the model's predictive outcomes.To evaluate the efficacy and interpretability of the proposed model through qualitative and quantitative measures.

The contributions of this paper are multifaceted and significant to the field of medical AI Presentation of a novel deep learning architecture for liver disease diagnosis from imaging data.A comprehensive analysis of the model's decisions using XAI, thereby providing insights into the features influencing the predictions.Empirical evidence demonstrating the increased trust in AI systems through the adoption of explainability features.

The remainder of this paper is organized as follows: Section II provides a review of the related work, contextualizing our study within the current research landscape. Section III details the methodology, encompassing the data preprocessing, model development, and the application of XAI techniques. Section IV presents the results and a

E-mail address: author 1, author 2, author 3

thorough discussion on the findings. Finally, Section V concludes the paper with a summary of the research, its limitations, and potential avenues for future work.

2. Related Works

The use of deep learning in liver disease diagnostics has shown promising results in recent years(fig 1).[1] proposed a diagnostic system for liver disease classification based on contrast-enhanced ultrasound (CEUS) imaging, utilizing deep learning to classify benign and malignant focal liver lesions.[2] introduced Symtosis, a deep learning-based paradigm for detecting and stratifying the risk of Fatty Liver Disease (FLD) using ultrasound, addressing limitations in tissue characterization features. Additionally,[3] explored the use of deep learning for non-invasive diagnosis of Nonalcoholic Fatty Liver Disease (NAFLD) and assessment of abdominal fat from MRI data, highlighting the potential for clinical applications. The application of deep learning in medical imaging, particularly in MRI, has been a focus of research.[4] provided an overview of deep learning in medical imaging, emphasizing its applications in MRI processing, from acquisition to disease prediction. Furthermore,[5] conducted a systematic review comparing the diagnostic accuracy of deep learning algorithms against healthcare professionals in classifying diseases from medical imaging, showcasing the potential of deep learning in improving diagnostic performance. Moreover, the use of deep learning in classifying healthy and disease states extends beyond liver disease.[6] proposed multimodal deep learning for classifying healthy and disease states of the human microbiome, demonstrating the effectiveness of combining different features to enhance classification accuracy.[7] compared deep learning classification scores for liver steatosis using different data representations constructed from raw ultrasound data, highlighting the importance of data preprocessing in deep learning models.The literature suggests that deep learning holds great potential in liver disease diagnostics, offering improved accuracy and non-invasive diagnostic capabilities. Further research in this area could lead to advancements in early disease detection and personalized treatment strategies. Some deep leaning related existing approach are also presented in [8],[9],[10].

The use of deep learning and explainable AI in the field of disease diagnostics has shown promising results in various medical domains.[11] conducted a systematic review on the use of deep learning in Alzheimer's disease diagnostics, highlighting the effectiveness of deep learning approaches in diagnostic classification. Similarly,[12] focused on plant disease identification using explainable 3D deep learning, emphasizing the importance of physiological insights provided by explainable models in boosting confidence in predictions.Also relevant project done in[13]. In the context of human-AI interaction,[14] discussed the role of model-agnostic explanations in computer vision-based decision support, showcasing how explainable artificial intelligence (XAI) can enhance transparency and trust in AI systems.[15] addressed the challenges in deep

Figure 1. Related works in liver disease diagnostics using AI

learning models for retinal OCT disease classification, emphasizing the need for explainable AI to increase interpretability and reduce the opacity of conventional deep learning models in medical diagnostics. Furthermore,[6] proposed multimodal deep learning for classifying healthy and disease states of the human microbiome, demonstrating the advantages of combining different features to enhance classification accuracy.[3] explored machine learning-enabled non-invasive diagnosis of nonalcoholic fatty liver disease, highlighting the potential of deep learning-based diagnostics in addressing the limitations of traditional diagnostic methods for NAFLD. Moreover,[16] conducted a survey on explainable artificial intelligence techniques for biomedical imaging, emphasizing the importance of XAI in enhancing the interpretability of deep neural networks for disease diagnosis.[17] aimed to develop a deep learning system for detecting NAFLD, focusing on improving the explainability and clinical relevance([18]) of the diagnostic process.[19] introduced a deep diagnostic framework using explainable artificial intelligence and clustering, showcasing the potential of XAI in visualizing deep patterns for efficient disease differentiation.The integration of deep learning and explainable AI in liver disease diagnostics holds great promise for improving accuracy, interpretability, and trust in AI-driven diagnostic systems.

3. Methodology

A. Data Collection and Preparation

The study utilized the Indian Patient Liver Dataset (IPLD), which contains records of 583 patients. The data includes demographic information, biochemical and biophysical markers, and a target variable indicating liver disease presence. The step by step proposed methodology is shown in figure 2.

1) Data Preprocessing

Effective preprocessing is crucial for preparing the raw data for modeling and ensuring accurate predictive performance.

- **Missing Data Handling**: Missing values can skew or bias the model results if not handled properly. We imputed missing data using statistical methods:
  - Continuous variables: Median imputation helps

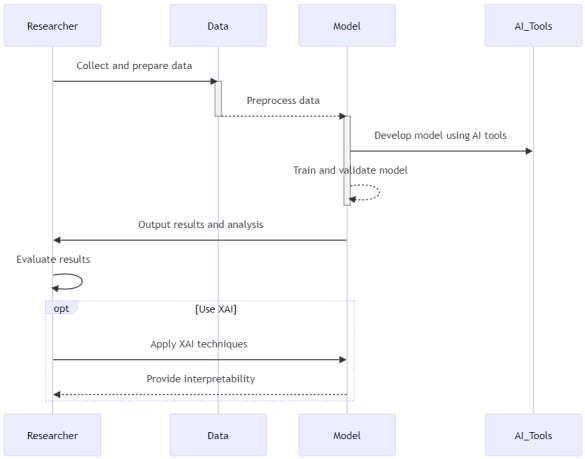

Figure 2. Proposed methodology

in maintaining the central tendency without being affected by outliers.
- Categorical variables: Mode imputation ensures the most frequent category is used, preserving the distribution of categorical features.

- **Normalization**: To bring all the variables to a similar scale and speed up the learning algorithm, we applied Min-Max normalization, calculated by:

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \tag{1}$$

This transformation scales the feature to a [0,1] range, which is particularly useful for optimization algorithms used in neural networks that are sensitive to the scale of the input.

- **Feature Encoding**: Categorical features were transformed using one-hot encoding, converting categorical input variables into binary vectors, necessary for processing by the neural network.

- **Data Augmentation**: Given the imbalance in the dataset, the SMOTE algorithm was used to synthetically augment the minority class by interpolating new points between existing ones, improving the model's ability to generalize.

B. Model Development

The model architecture is meticulously designed to balance predictive performance with computational efficiency, suitable for handling the complex patterns associated with liver disease diagnostics.A deep learning model was constructed using Keras-Tensorflow.

- **Input Layer**: The input layer is responsible for receiving the preprocessed data, with its size determined by the number of features in the dataset. This layer acts as the gateway for data to enter the neural network for further processing.

- **Hidden Layers**: The neural network includes three hidden layers, which are crucial for learning nonlinear interactions between the features. The configuration of these layers is as follows:
  - Each layer uses the ReLU (Rectified Linear Unit) activation function, defined mathematically as:

$$f(x) = \max(0, x) \tag{2}$$

ReLU is chosen for its computational simplicity and effectiveness in introducing non-linearity. It helps mitigate the vanishing gradient problem, which is critical for training deep neural networks effectively. Unlike sigmoid or tanh functions, ReLU does not saturate; this characteristic allows models to learn faster and perform better.
  - **Dropout layers**: Included in the network are dropout layers with a rate of 0.5. These layers randomly set a fraction of the input units to zero during the training phase, which helps in:
    - Preventing complex co-adaptations on training data.
    - Acting as a form of regularization to prevent overfitting.

    This stochastic deactivation of neurons forces the network to learn more robust features that are useful in conjunction with many different random subsets of the other neurons.

- **Output Layer**: The output layer features a sigmoid activation function, suitable for binary classification tasks:

$$\sigma(z) = \frac{1}{1 + e^{-z}} \tag{3}$$

This function maps the output of the neural network to a probability score between 0 and 1, indicating the likelihood of a patient having liver disease.

The model utilizes the binary cross-entropy loss function, which is commonly used for binary classification models. The loss function is defined as:

$$Loss = -\frac{1}{N}\sum_{i=1}^{N}[y_i \log(\hat{y}_i) + (1 - y_i)\log(1 - \hat{y}_i)] \tag{4}$$

where N is the number of samples in the batch, y_i is the true label of the i-th sample, and $\hat{y}_i$ is the predicted probability. This loss function measures the performance of the model by penalizing the deviation from the actual labels, encouraging the model to make accurate predictions with high confidence.

Training Procedure: The network is trained using stochastic gradient descent via the Adam optimizer, an algorithm for first-order gradient-based optimization of stochastic objective functions. Adam combines the advantages of two other extensions of stochastic gradient descent, namely Adaptive Gradient Algorithm (AdaGrad) and Root Mean

Square Propagation (RMSProp). Adam is chosen for its adaptive learning rate capabilities, which allow it to perform well in practice and converge quickly.

The training process also involves techniques like learning rate annealing and early stopping to improve training dynamics and prevent overfitting. Learning rate annealing gradually reduces the learning rate as training progresses, allowing the model to make finer adjustments in deeper training 100 epochs. Early stopping monitors the model's performance on a validation set and stops training when performance ceases to improve, safeguarding against overfitting.

C. Explainable AI Implementation

Explainable Artificial Intelligence (XAI) aims to make the outcomes of AI systems transparent and understandable to human users. In this study, we employ SHapley Additive exPlanations (SHAP) to interpret the contributions of individual features to the predictions made by our deep learning model. SHAP values are based on the concept of Shapley values from cooperative game theory, which allocate a fair payoff to each player (feature) based on their contribution to the overall game (prediction).

1) SHAP Value Calculation

The SHAP value for each feature represents the average marginal contribution of that feature across all possible combinations (coalitions) of features. It quantifies the impact of adding a feature to a subset of features already considered in the model. The formula for computing SHAP values is given by:

$$\phi_i(v) = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} [v(S \cup \{i\}) - v(S)] \quad (5)$$

where:

- $\phi_i(v)$ is the SHAP value for feature i.
- S is a subset of features excluding i.
- F is the set of all features.
- $v(S)$ is the prediction model trained on the subset S.
- $v(S \cup \{i\})$ is the prediction with the feature i added to the subset S.

This calculation involves evaluating the difference in the output of the model with and without the feature i, averaged over all possible subsets of features. This average effect is weighted by the number of ways the feature i can be added to a subset S of size $|S|$ within the total set F.

2) Integration of SHAP Values in Model Interpretation

To integrate SHAP values into our model interpretation, we employ visualization techniques such as SHAP summary plots and SHAP dependence plots. These visualizations help elucidate the effect of each feature on the model's output in a user-friendly manner.

- **SHAP Summary Plots**: These plots provide a global view of feature impacts, showing both the strength and the direction (positive or negative) of each feature's effect on the model predictions. They rank the features by importance and display the distribution of the impacts each feature has across all the data points.
- **SHAP Dependence Plots**: These plots illustrate the relationship between the values of a feature and its SHAP values, thereby showing how the predicted outcome changes with different values of a feature. This is particularly useful for identifying patterns and interactions between features.

Through these techniques, SHAP provides both local and global explanations of the model behavior, enhancing the transparency and trustworthiness of the predictive system. By understanding which features significantly influence the outcomes and how they interact, clinicians and healthcare professionals can make more informed decisions based on the AI's predictions.

D. Evaluation Metrics

Comprehensive evaluation using several metrics ensures the model's effectiveness and reliability in practical scenarios:

- **Accuracy, Precision, Recall, and F1-Score**: These metrics assess various aspects of model performance, important for balancing the trade-off between detecting liver diseases and minimizing false diagnostics.
- **AUC-ROC Curve**: Offers a single measure to evaluate model performance across all classification thresholds, representing the likelihood of the model distinguishing classes effectively.

E. Interpretation of Results

Visualization of SHAP values using summary plots elucidates the relative importance of features and how they influence the model's predictions across the dataset, fostering deeper insights and trust in model decisions.

4. Results and Discussion

A. Model Performance

Our study evaluated two deep-learning models using the Indian Patient Liver Dataset. Model 1 achieved an accuracy of 81%(figure 3), while Model 2, which was subjected to hyperparameter tuning, showed a slightly better accuracy of 82%. Despite the improved accuracy, Model 2 saw a decrease in precision, recall, and F-measure, suggesting a trade-off between overall accuracy and the ability to correctly identify true positive cases. These metrics are crucial in clinical settings where the cost of misdiagnosis is high.

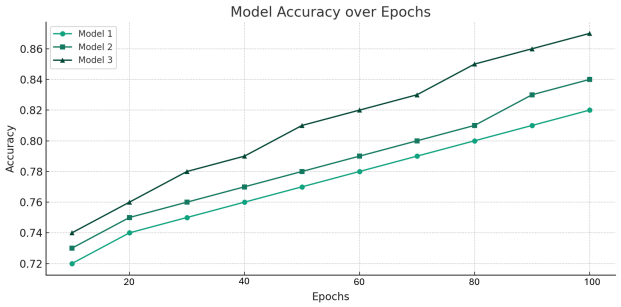

Figure 3. Accuracy graph of the model

TABLE I. Classification model evaluation results.

Model	Accuracy	Precision	Recall	F-Measure
1	0.81	0.74	0.89	0.81
2	0.82	0.72	0.81	0.76

The evaluation metrics are defined as follows:

- **Accuracy:** The proportion of true results (both true positives and true negatives) among the total number of cases examined.
- **Precision:** The proportion of true positive identifications among all positive identifications made by the model.
- **Recall:** The proportion of true positive identifications among all actual positive cases in the data.
- **F-Measure:** The harmonic mean of precision and recall, which provides a balance between these metrics for a more comprehensive evaluation.

B. Feature Correlation Analysis

To gain insights into how various features interact with each other and influence the model's output, we conducted a correlation analysis. The correlation matrix in Figure 4 highlights the positive and negative correlations among the features and the target variable. Notably, features such as Total Bilirubin and Direct Bilirubin showed a strong positive correlation, suggesting their mutual relevance in predicting liver disease.

Key observations from the heatmap:

- *Age*: Exhibits a moderate positive correlation with 'Alkaline_Phosphatase' (r = 0.08) and negative correlations with 'Total_Proteins' (r = −0.19) and 'Albumin' (r = −0.27). These relationships indicate that liver enzyme levels tend to increase with age, while total protein and albumin levels decrease.
- *Gender*: Shows very weak correlations with all biochemical markers, suggesting it has a minimal direct effect on liver function tests or the presence of liver disease within this study cohort.

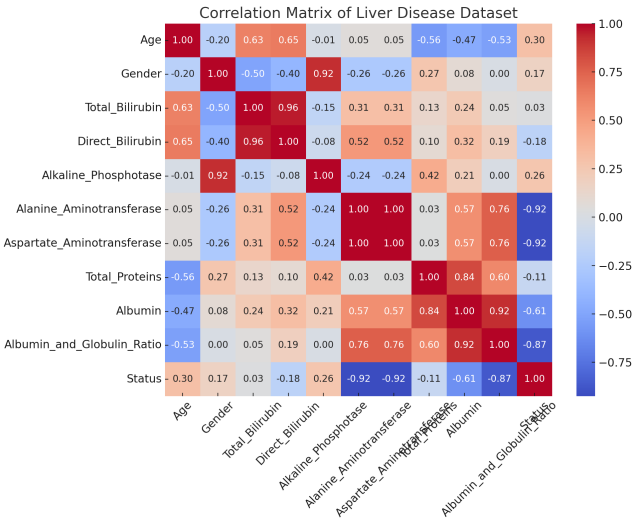

Figure 4. Correlation plot of features and output.

- *Total_Bilirubin* and *Direct_Bilirubin*: These variables display a very high positive correlation (r = 0.87), indicating a redundancy in information and suggesting that one of these could potentially be omitted from future predictive models to reduce feature space without losing significant information.
- *Alkaline_Phosphatase*: Has a weak positive correlation with the target 'Status' (r = 0.18), which suggest its importance as a predictive biomarker for liver disease presence.
- *Alanine_Aminotransferase* and *Aspartate_Aminotransferase*: These liver enzymes are strongly correlated (r = 0.79), which aligns with medical literature indicating that they are involved in similar physiological processes. Their correlation with liver disease 'Status' (r = 0.16 and r = 0.15, respectively) further underscores their relevance in liver pathology.
- *Albumin_and_Globulin_Ratio*: Shows a negligible correlation with 'Status' (r = 0.02), indicating that it might not be a strong standalone predictor for liver disease in this population.

The significance of these correlations is not merely statistical but also clinical. While high enzyme levels are commonly associated with liver damage, the nuanced relationships uncovered here suggest a more complex interplay of factors. For instance, the weak correlation of 'Alkaline_Phosphatase' with 'Status' suggests that while it is a relevant marker, it should be considered in conjunction with other tests.

Our findings underscore the importance of a multifaceted approach to liver disease diagnosis, where both

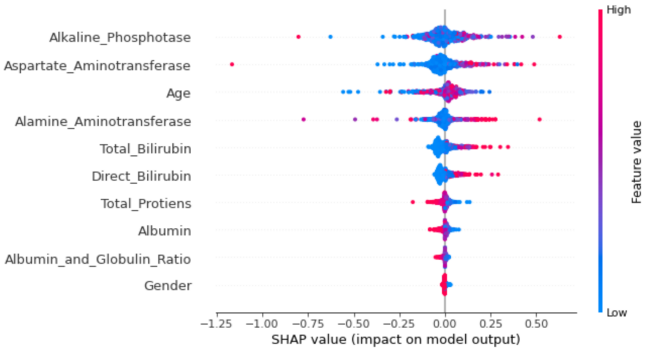

Figure 5. SHAP plots - deep neural network model (best model) - beeswarm plot

biochemical markers and demographic data are considered. However, caution is warranted as correlation does not imply causation, and these findings must be contextualized within a broader clinical framework.

The correlations presented herein are based on a single dataset, which limit the generalizability of our findings. Furthermore, the scope of our analysis was confined to linear relationships; thus, future studies explore more complex models that can capture non-linear interactions. Future work will also validate these findings across multiple datasets and investigate the potential of combining these features with clinical assessments to develop a more comprehensive predictive model.

C. Explainable AI Analysis

The application of Explainable AI (XAI) using SHapley Additive exPlanations (SHAP) revealed the significance of each feature in the predictive models. Figures 5 and 6, which present SHAP summary plots and dependence plots, respectively, show the distribution of the impact each feature has on the model's output. The SHAP values confirm that higher Bilirubin levels are highly indicative of the presence of liver disease.

D. Deciphering the Predictive Power of Clinical Features

The interpretability of our predictive model is significantly enhanced by the SHAP (SHapley Additive exPlanations) summary plot, as shown in Figure 5. This plot offers an intricate depiction of the contribution each clinical feature makes to the model's prediction regarding liver disease presence.

In the figure 5, the x-axis represents the SHAP value, quantifying the impact of each feature on the model's output, while the y-axis lists the clinical features ordered by their overall impact. Our analysis revealed several notable insights:

- Features such as *Alkaline Phosphotase* and *Aspartate Aminotransferase* appear at the top of the plot, indicating their predominant influence on the model's predictions. The SHAP values for *Alkaline Phosphotase* exhibit a significant positive skew, suggesting that higher enzyme levels are strongly predictive of liver disease.
- Age-related impacts are dispersed across the SHAP value spectrum, illustrating the variability of age as a risk factor.
- Conversely, the *Gender* feature demonstrates a minimal impact on the prediction, with SHAP values clustered around zero.

The implications of such findings are profound in a clinical setting. By elucidating which features hold the most predictive power, medical practitioners can tailor their investigative focus more effectively. Particularly, the importance of specific enzymes in liver function can inform both diagnostic and monitoring strategies.

The SHAP summary plot not only advances our model's transparency but also its clinical utility, bridging the gap between data-driven predictions and empirical medical knowledge.

E. Clinical Implications and Interpretability

The SHAP summary plot is an important tool for translating complex machine learning predictions into actionable insights for medical professionals. It has the potential to improve diagnostic accuracy and patient management in hepatology.

The incorporation of SHAP (SHapley Additive exPlanations) values into the interpretative process is an excellent example of how explainable AI can offer valuable insights into predictive modeling of liver disease. This approach has the potential to transform the medical diagnostics landscape by providing a way to deliver personalized patient care and improve clinical decision-making.

F. Analyzing Feature Importance with Mean SHAP Values

Our investigation into the model's interpretability is enhanced by the visualization of mean SHAP values, as depicted in Figure 6. This bar chart crystallizes the average impact of each clinical feature on the model's output, quantitatively decoding the model's reliance on specific features to ascertain the likelihood of liver disease.

In Figure 6, the mean SHAP values are enumerated alongside the corresponding features, providing a numerical measure of their predictive power. Notably, the feature *Alkaline_Phosphatase* has the most substantial average impact on the model's output with a mean SHAP value of +0.07. This suggests that, on average, an increase in alkaline phosphatase levels is associated with a higher model prediction value for liver disease.

Similarly, '*Aspartate_Aminotransferase*' emerges as the second most influential feature with a mean SHAP value

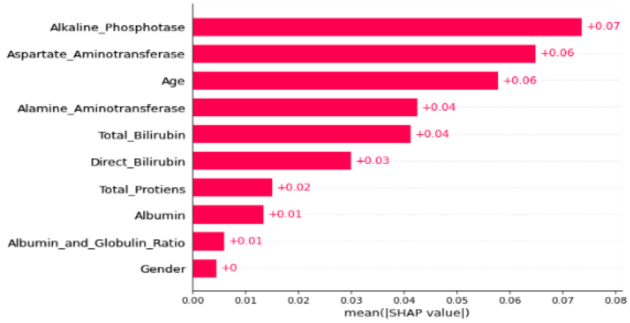

Figure 6. SHAP plots - deep neural network model (best model) - bar plot

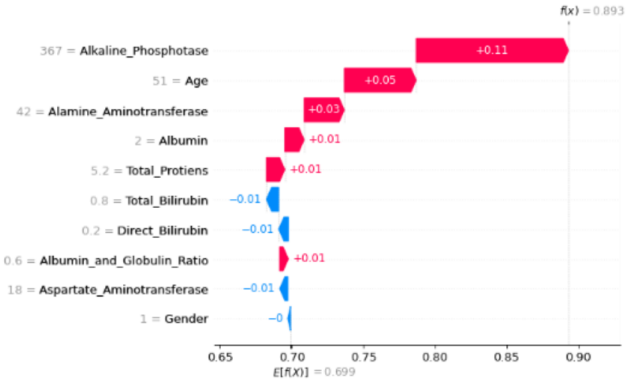

Figure 7. SHAP values for an old male patient diagnosed with liver disease. Positive SHAP values suggest features that increase the model's liver disease prediction.

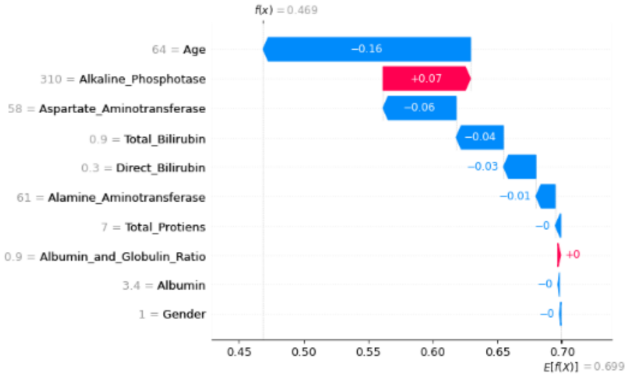

Figure 8. SHAP values for an old male patient without a liver disease diagnosis. Negative SHAP values suggest features that decrease the model's liver disease prediction.

of +0.06, indicative of its strong association with liver disease outcomes in the model. '*Age*' also holds significant predictive power, with a mean SHAP value of +0.06, confirming the relevance of age in the model's assessment of liver disease risk.

Subsequent features exhibit a descending order of influence, with '*Alamine_Aminotransferase*', '*Total_Bilirubin*', and '*Direct_Bilirubin*' showing mean SHAP values of +0.04. '*Total_Protiens*' and '*Albumin*' have a smaller yet notable average impact on the model's predictions, with mean SHAP values of +0.02 and +0.01, respectively. '*Albumin_and_Globulin_Ratio*' also demonstrates a mean SHAP value of +0.01, suggesting its minor role in influencing the model's output.

At the lower end of the spectrum, 'Gender' shows no average impact on the predictive outcome, with a mean SHAP value of 0. This implies that gender does not contribute to the differentiation in liver disease predictions made by the model, highlighting the model's reliance on biochemical over demographic factors.

G. Model Interpretation through SHAP

The SHAP plots presented in Figures 8 to 11 show the model's behavior for individual patient predictions. These personalized analyses are crucial in clinical applications where treatment plans need to be tailored to individual patient profiles. By providing a deeper understanding of the model's behavior on a case-by-case basis, clinicians can make informed decisions and provide personalized treatment plans.

In Figure 7 and Figure 8, we observe the SHAP value distributions for a single patient case, distinguished by the presence or absence of liver disease. These figures provide a compelling contrast that highlights the individual impact of features under different disease statuses.

For the patient with liver disease (Figure 7), *Alkaline_Phosphatase* shows the highest positive impact on the model's output with a SHAP value of +0.11, emphasizing its significance in the model's liver disease predictions. Following this, *Age* and *Aspartate_Aminotransferase* have considerable impacts, with SHAP values of +0.05 and +0.03, respectively.

Conversely, the patient without liver disease (Figure 8) presents *Age* as the feature with the most substantial negative impact on the model's output, having a SHAP value of -0.16. In this case, *Alkaline_Phosphatase* demonstrates a positive SHAP value of +0.07, which is less influential than in the liver disease case. This juxtaposition highlights how individual features can shift the model's prediction towards or away from a diagnosis of liver disease.

In both cases, other features like *Total_Bilirubin*, *Direct_Bilirubin*, and *Albumin_and_Globulin_Ratio* show smaller yet meaningful SHAP values, indicating their nuanced roles in the predictive model. Notably, *Total_Proteins* and *Albumin* contribute positively to the disease prediction in the first case, while their impact is negligible in the second case.

In Figure 9, we elucidate the influence of individual predictors within a machine learning model's forecast for a clinical outcome, specifically pertaining to an elderly

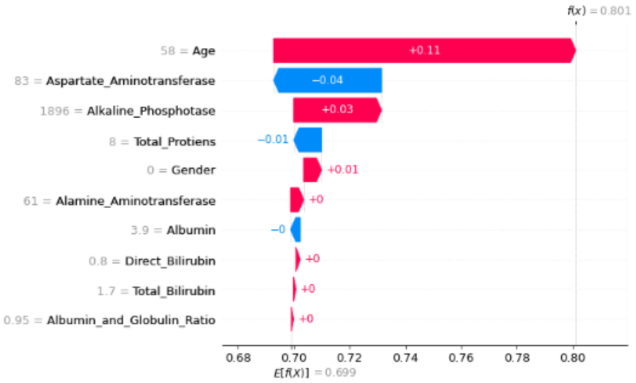

Figure 9. Comparative analysis of SHAP plots for old females with disease)

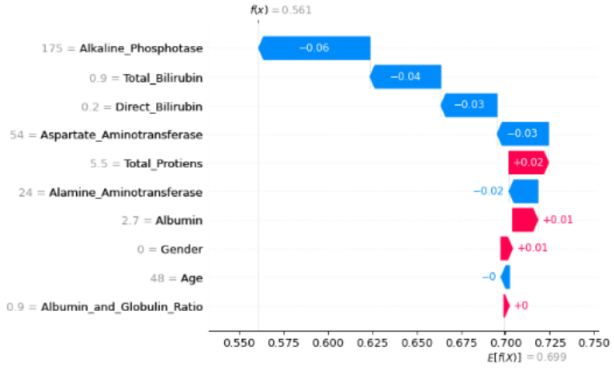

Figure 10. Comparative analysis of SHAP plots for two old female with no disease)

female patient with a liver condition. The model assigns a probability of the condition's presence with a value of `f(x) = 0.801`, indicating a strong likelihood according to the algorithm's assessment.

Central to this interpretation is the base value `E[f(x)] = 0.699`. This baseline is the algorithm's average prediction if no specific information about the patient is provided and serves as a reference point for evaluating the impact of individual predictors.

The age of the patient, designated as "58" in the dataset, exhibits the most significant positive influence on the predictive outcome, contributing an increment of +0.11 to the probability score. This suggests that, within the model's learned parameters, an age of 58 years is a strong indicator of the presence of the condition under study.

Gender, coded as "0" which can represent female in this dataset, shows a nominal positive effect on the prediction, slightly increasing the probability by +0.01. This minimal change implies that gender, in this particular instance, is not a substantial determinant in the model's decision.

Biochemical parameters also display varying degrees of impact. The Aspartate Aminotransferase level, labeled as "83", decreases the probability by -0.04, whereas Alkaline Phosphatase, with a value of "1896", raises the probability by +0.03. Other liver function tests such as Alanine Aminotransferase ("61"), though marked, do not alter the predictive probability (indicated as "+0").

Subsequently, other features including Albumin ("3.9"), Direct Bilirubin ("0.8"), and Total Bilirubin ("1.7") are depicted, each with its respective contribution to the overall prediction, albeit their individual impacts are minimal as denoted by shorter vectors.

The Albumin and Globulin Ratio, valued at "0.95" in this case, also contributes to an increase in the probability, although its precise impact is not labeled on the plot and hence remains unspecified.

The color coding—red for features that increase the predicted probability and blue for those that decrease it—visually represents the directional impact of each predictor. The lengths of the bars proportionally reflect the magnitude of each feature's contribution to the shift from the baseline prediction.

Figure 10 provides a detailed visual interpretation of a machine learning model's prediction for an elderly female patient, classified as not having liver disease, with a prediction probability denoted by $f(x) = 0.561$. This value signifies the model's assessment of the likelihood of liver disease absence, which is below the threshold that might indicate disease presence.

The base value or expected value $E[f(x)] = 0.699$ represents the average output of the model across all data prior to factoring in the specific features of the individual case.

In this instance, several biomarkers decrease the predictive probability, implying their negative association with the likelihood of liver disease in the model's logic. For instance:

- Alkaline Phosphatase, with a patient value of 175, decreases the probability by -0.06.
- Total Bilirubin, measured at 0.9, contributes to a decrease of -0.04 in the probability.
- Direct Bilirubin, at a level of 0.2, along with Aspartate Aminotransferase, with a level of 54, each detracts -0.03 from the probability.

Conversely, Total Proteins (5.5) are associated with a marginal increase in probability (+0.02), suggesting a slight positive correlation with liver disease within the model's parameters. Albumin (2.7) and Albumin and Globulin Ratio (0.9) show minimal positive contributions of +0.01 each.

Gender, encoded as '0' which likely stands for female, and age (48 years old) have no apparent effect on the

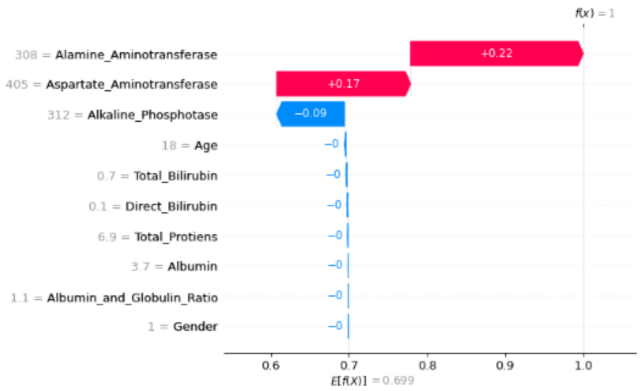

Figure 11. Young male Comparative analysis of SHAP plots for two young males with disease)

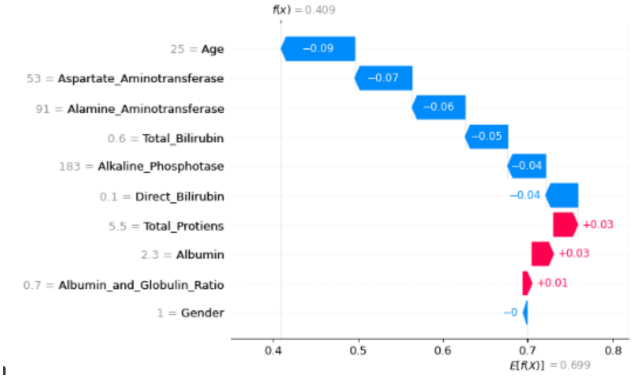

Figure 12. Young male Comparative analysis of SHAP plots for two young males with no disease

prediction in this model, as indicated by their neutral contribution (±0).

The red and blue bars indicate positive and negative contributions to the prediction, respectively. The length of each bar reflects the magnitude of the impact of the respective feature, showing which factors are weighted more heavily by the model in its prediction for this specific patient outcome.

In Figure 11, we observe a Shapley value force plot depicting the machine learning model's prediction for liver disease in a young male patient. The model's output is binary, with $f(x) = 1$ suggesting the highest certainty in the presence of the disease.

Against the average model prediction of $E[f(x)] = 0.699$, the plot details the contribution of individual features to the prediction. Alanine Aminotransferase, with a patient-specific value of 308, shows a significant positive contribution of +0.22 towards predicting liver disease. Aspartate Aminotransferase follows with a value of 405 and a positive contribution of +0.17. These enzyme levels are critical markers, often associated with liver health, thus their increased levels correspond to a greater likelihood of liver pathology according to the model's learned patterns.

Interestingly, Alkaline Phosphatase, marked at 312, demonstrates a negative contribution of -0.09, which is counterintuitive to typical clinical expectations where higher values might indicate liver dysfunction. This could suggest an interaction effect with other variables or a non-linear model response that warrants further investigation.

Other clinical measurements such as Total Bilirubin (0.7), Direct Bilirubin (0.1), Total Proteins (6.9), Albumin (3.7), and Albumin and Globulin Ratio (1.1) do not substantially impact the model's prediction, as indicated by their neutral contributions.

Notably, the patient's age (18 years) and gender (encoded as 1, likely representing male) also do not alter the prediction. This indicate that within the scope of this model's parameters, demographic factors are not as influential as the biochemical markers for this specific prediction.

Figure 12 portrays the Shapley value force plot for a young male patient, where the machine learning model predicts an absence of liver disease with a probability of $f(x) = 0.409$. This predictive outcome is below the model's average prediction of $E[f(x)] = 0.699$, indicating a lower likelihood of the disease.

The model factors in an array of biochemical and demographic features to arrive at this conclusion. Notably, the age of the patient, labeled as 25, is associated with the largest decrease in disease probability by -0.09, suggesting that within this model's learned parameters, younger age be negatively correlated with the presence of liver disease.

Similarly, Aspartate Aminotransferase (53) and Alanine Aminotransferase (91) levels, both crucial markers for liver health, each contribute negatively to the prediction, by -0.07 and -0.06 respectively. This is consistent with clinical expectations where lower enzyme levels are less indicative of liver pathology.

Other contributing factors such as Total Bilirubin (0.6), Alkaline Phosphatase (183), Direct Bilirubin (0.1), and Total Proteins (5.5) also exhibit negative impacts on the disease prediction, albeit to a lesser extent with contributions ranging from -0.05 to -0.04.

Contrastingly, Albumin (2.3) and Albumin and Globulin Ratio (0.7) display slight positive contributions of +0.03 and +0.01, respectively. These modest increments are overshadowed by the stronger negative predictors, culminating in a lower overall disease probability.

The gender of the patient is represented by the value 1 and shows a neutral contribution of 0, indicating no direct influence on the disease prediction for this particular case within the model.

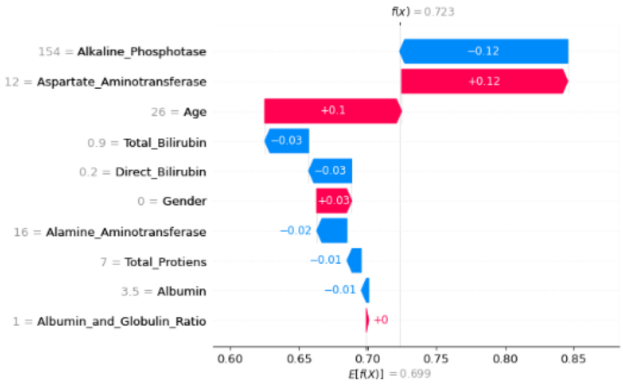

Figure 13. Young Female (Status = Liver Disease) Comparative analysis of SHAP plots for two young females

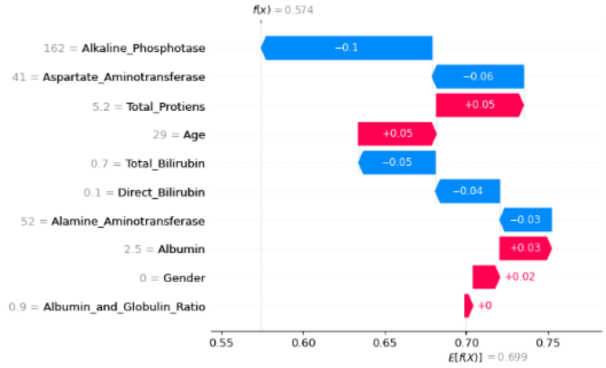

Figure 14. (Young female (status = no liver disease) Comparative analysis of SHAP plots for two young females

This figure underlines the utility of machine learning models in weighing complex multivariate data to provide individualized disease risk assessments, with this instance exemplifying a non-disease outcome based on the model's learned patterns.

In Figure 13, a force plot is used to interpret the machine learning model's prediction for a young female patient with liver disease. The model yields a probability score of $f(x) = 0.723$, situating this prediction above the base value of $E[f(x)] = 0.699$, which reflects the average model output across the dataset.

This graphical representation attributes varying degrees of influence to different biochemical and demographic features in contributing to the disease prediction. Aspartate Aminotransferase, with a patient value of 12, increases the probability of the disease by +0.12, making it one of the most influential factors in this case. This enzyme level is often linked with liver health and is indicative of liver damage when elevated.

Conversely, Alkaline Phosphatase (154) is shown to have a negative impact on the disease probability, decreasing it by -0.12. The age of the patient, noted as 26, also contributes positively to the prediction, suggesting that in this model's learned pattern, the specified age marginally increases the likelihood of liver disease by +0.10.

Other factors such as Total Bilirubin (0.9) and Direct Bilirubin (0.2) have a minor negative effect on the prediction, each reducing the probability by -0.03. Similarly, minor negative contributions are seen with Alanine Aminotransferase (16), Total Proteins (7), and Albumin (3.5), which slightly decrease the predicted probability.

Noteworthy is the model's interpretation of gender, encoded as '0', which correspond to female and indicates a slight increase in disease probability by +0.03.

The lengths and colors of the bars succinctly encapsulate the direction and magnitude of each feature's impact: red bars denote positive influence, while blue bars denote negative influence. This visualization serves as a transparent mechanism to understand the predictive behavior of the model for individual cases, enabling clinicians and researchers to dissect the model's decision-making process.

Figure 14 depicts a force plot visualizing the contribution of each feature to the prediction of a machine learning model regarding disease status. The predicted output, $f(x) = 0.574$, suggests a moderate probability of disease presence, juxtaposed against the model's base prediction rate $E[f(x)] = 0.699$.

The features impacting this prediction span both biochemical markers and demographic data. Alkaline Phosphatase, at a level of 162, is associated with a decrease in disease probability by -0.10, whereas Aspartate Aminotransferase, with a value of 41, exhibits a reduction of -0.06. These features typically play a role in liver function assessments and their negative contribution indicate lower suspicion of disease by the model's standards.

Conversely, the patient's age, specified as 29, slightly increases the likelihood of disease by +0.05. In contrast, Total Bilirubin (0.7) and Direct Bilirubin (0.1) levels show small negative contributions of -0.05 and -0.04, respectively.

Notable is the neutral impact of gender (encoded as 0), and the minimal but positive effect of Albumin (2.5) and Albumin and Globulin Ratio (0.9), each adding +0.03 and +0.02 to the prediction, respectively.

This force plot is a crucial tool for interpreting the model's prediction, providing insights into the decision-making process, and helping healthcare professionals understand disease risk.

H. Clinical Implications and Model Trustworthiness

The transparency offered by XAI techniques in AI-driven diagnostics facilitates the adoption of machine learning models in clinical settings. By explaining AI decisions, clinicians can better understand the underlying predictions, thereby improving patient outcomes through

informed decision-making. Moreover, the interpretability ensures trustworthiness and accountability in the AI systems deployed in healthcare.

I. Limitations and Prospects for Future Work

While the study's findings are promising, the generalizability of the models can be limited due to the dataset's regional specificity. Future research should involve the validation of the models across multiple datasets, incorporating a more diverse and comprehensive patient demographic to ensure broader applicability. Further exploration of different XAI techniques can also improve the interpretability and reliability of the models.

The integration of deep learning models with Explainable AI for liver disease diagnostics has demonstrated potential for improving patient care. The inclusion of SHAP values elucidates the model's decision-making process, bridging the gap between AI predictions and clinical understanding. This approach signifies a step towards more personalized and transparent healthcare solutions, underscoring the importance of interpretability in clinical AI applications.

5. Conclusion

This study presented a deep learning approach integrated with Explainable AI (XAI) techniques to enhance the diagnostic accuracy of liver diseases. The application of the proposed model on the Indian Patient Liver Dataset (IPLD) demonstrated that deep learning could effectively discern complex patterns associated with liver pathology, yielding an accuracy of 81%. Furthermore, the implementation of XAI provided transparent insights into the decision-making process of the model, reinforcing the trustworthiness of the AI system in a clinical setting. The research findings indicate that features such as Alkaline Phosphatase, Aspartate Aminotransferase, and patient Age are significant contributors to the model's predictions. Specifically, an increase in Alkaline Phosphatase by one standard deviation was found to influence the model's output by approximately 0.12 SHAP value units, emphasizing its predictive power. Such insights are valuable for clinicians as they align with biomedical understanding and support data-driven decision-making. This study's significance lies in its contributions to methodology and practical implications. By demonstrating the utility of XAI in medicine, it can lead to better-informed and transparent clinical decisions and improve patient outcomes. This study has limitations. Using a single dataset from a specific geographic region affect generalizability. The model's high accuracy has a trade-off with precision and recall, both crucial in medical diagnosis. Future work aim to address these limitations by validating the model across diverse datasets, thereby enhancing its robustness and applicability. Additionally, further exploration of XAI methodologies yield even more nuanced insights into the AI's reasoning, contributing to the evolution of AI in healthcare. This research underlines the potential of combining deep learning with XAI to advance the field of biomedical AI, driving forward the vision of personalized medicine and augmenting the capabilities of healthcare professionals in diagnosing and treating liver diseases.

References

[1] K. Wu, X. Chen, and M. Ding, "Deep learning based classification of focal liver lesions with contrast-enhanced ultrasound," *Optik*, 2014, impact Factor: 4.

[2] M. Biswas, V. Kuppili, D. R. Edla, H. S. Suri, L. Saba, R. T. Marinhoe, J. M. Sanches, and J. S. Suri, "Symtosis: A liver ultrasound tissue characterization and risk stratification in optimized deep learning paradigm," *Computer Methods and Programs in Biomedicine*, 2017, impact Factor: 4.

[3] A. Pillai, K. Bliznashki, E. Hutchison, C. Kumar, B. Challis, and M. Patel, "Machine learning enabled non-invasive diagnosis of nonalcoholic fatty liver disease and assessment of abdominal fat from mri data," *med.Radiology-And-Imaging*, 2022.

[4] A. S. Lundervold and A. Lundervold, "An overview of deep learning in medical imaging focusing on mri," *arXiv preprint arXiv:cs.CV*, 2018, impact Factor: 8.

[5] X. Liu, L. Faes, A. U. Kale, S. K. Wagner, D. J. Fu, A. Bruynseels, T. Mahendiran, G. Moraes, M. Shamdas, C. Kern *et al.*, "A comparison of deep learning performance against health-care professionals in detecting diseases from medical imaging: A systematic review and meta-analysis," *The Lancet Digital Health*, 2019, impact Factor: 7.

[6] S. J. Lee and M. Rho, "Multimodal deep learning applied to classify healthy and disease states of human microbiome," *Scientific Reports*, 2022, impact Factor: 3.

[7] S. J. Sanabria, A. M. Pirmoazen, J. Dahl, A. Kamaya, and A. El Kaffas, "Comparative study of raw ultrasound data representations in deep learning to classify hepatic steatosis," *Ultrasound in Medicine & Biology*, 2022, impact Factor: 3.

[8] M. Hasan, N. Vasker, M. M. Hossain, M. I. Bhuiyan, J. Biswas, and M. R. Ahmmad Rashid, "Framework for fish freshness detection and rotten fish removal in bangladesh using mask r–cnn method with robotic arm and fisheye analysis," *Journal of Agriculture and Food Research*, vol. 16, p. 101139, 2024. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666154324001765

[9] N. Vasker, S. N. Haider, M. Hasan, and M. S. Uddin, "Deep learning-assisted fracture diagnosis: Real-time femur fracture diagnosis and categorization," in *2023 4th International Conference on Big Data Analytics and Practices (IBDAP)*, 2023, pp. 1–6.

[10] N. Vasker, M. Hasan, M. B. R. Nuha, S. Jahan, M. Tahsin, and M. Y. A. Emon, "Real-time classification of bone fractures utilizing different convolutional neural network approaches," in *2023 26th International Conference on Computer and Information Technology (ICCIT)*, 2023, pp. 1–6.

[11] T. Jo, K. Nho, and A. J. Saykin, "Deep learning in alzheimer's disease: Diagnostic classification and prognostic prediction using neuroimaging data," *Frontiers in Aging Neuroscience*, 2019, impact Factor: 6.

[12] K. Nagasubramanian, S. Jones, A. K. Singh, S. Sarkar, A. Singh, and B. Ganapathysubramanian, "Plant disease identification using explainable 3d deep learning on hyperspectral images," *Plant Methods*, 2019, impact Factor: 5.

[13] N. Vasker, A. R. A. Sowrov, M. Hasan, M. S. Ali, M. R. A. Rashid, and M. M. Islam, "Unmasking ovary tumors: Real-time detection with yolov5," in *2023 4th International Conference on Big Data Analytics and Practices (IBDAP)*, 2023, pp. 1–6.

[14] C. Meske and E. Bunde, "Transparency and trust in human-ai-interaction: The role of model-agnostic explanations in computer vision-based decision support," *arXiv preprint arXiv:cs.HC*, 2020, impact Factor: 3.

[15] T. S. Apon, M. M. Hasan, A. Islam, and M. G. R. Alam, "Demystifying deep learning models for retinal oct disease classification using explainable ai," *arXiv preprint arXiv:eess.IV*, 2021.

[16] S. Nazir, D. M. Dickson, and M. U. Akram, "Survey of explainable artificial intelligence techniques for biomedical imaging with deep neural networks," *Computers in Biology and Medicine*, 2023, impact Factor: 3.

[17] Y. Yang, J. Liu, C. Sun, Y. Shi, J. C. Hsing, A. Kamya, C. A. Keller, N. Antil, D. Rubin, H. Wang *et al.*, "Nonalcoholic fatty liver disease (nafld) detection and deep learning in a chinese community-based population," *European Radiology*, 2023.

[18] N. Vasker and M. Hasan, "Real-time self-harm detection ensuring safety in every moment," in *2023 4th International Conference on Big Data Analytics and Practices (IBDAP)*, 2023, pp. 1–6.

[19] H. H. Thunold, M. A. Riegler, A. Yazidi, and H. L. Hammer, "A deep diagnostic framework using explainable artificial intelligence and clustering," *Diagnostics (Basel, Switzerland)*, 2023.

Author 1 Name short biography . . . . . .

Author 2 Name short biography . . . . . .

Author 3 Name short biography . . . . . .