https://doi.org/10.31449/inf.v48i1.4611

Liver Disease Classification - An XAI Approach to Biomedical AI

Ebenezer Agbozo, Daniel Musafiri Balungu
Ural Federal University, Russian Federation
E-mail: eagbozo@urfu.ru, danielbal03.db@gmail.com

Keywords: deep learning, biomedical science, explainable ai, data-driven decision making, shapley values, prescriptive analytics

Received: Januar 12, 2023

Explosive amounts of biological and physiological data, including medical images, electroencephalograms, genomic information, and protein sequences, have been made available to us thanks to advances in biological and medical technologies. Understanding human health and disease is made easier by using this data for learning. Deep learning-based algorithms, which were developed from artificial neural networks, have significant potential for identifying patterns and extracting features from large amounts of complex data. However, these recent advancements involve blackbox models: algorithms that do not provide human-understandable explanations in support of their decisions. This limitation hampers the fairness, accountability and transparency of these models; the field of XAI tries to solve this problem providing human-understandable explanations for black-box models. This paper focuses on the requirement for XAI to be able to explain in detail the decisions made by an AI in a biomedical setting to the expert in the domain, e.g., the physician in the case of AI-based clinical decisions related to diagnosis, treatment, or prognosis of a disease. In this paper, we made use of the Indian Patient Liver Dataset (IPLD) collected from Andhra Pradesh region. The deep learning model with a 0.81 accuracy score (0.82 for the hyperparameter- tuned model) is built on Keras-Tensorflow and due to the imbalance in the target values, we integrated GANs as a means of oversampling the dataset. This study integrated the XAI concept of Shapley Values to shed light on the predictive results obtained by the liver disease detection model.

Povzetek: Študija obravnava klasifikacijo jetrnih bolezni z uporabo razložljive umetne inteligence (XAI), ki omogoča razumevanje odločitev modelov globokega učenja z integracijo Shapley vrednosti za razlago prediktivnih rezultatov.

1 Introduction

For most of its history, medicine was practiced on artistic principles rather than according to modern definitions of science. In the past two centuries, the practice of medicine has been more closely aligned with scientific method principles, particularly in regards to comprehending the molecular causes of disease. Advances in anatomy, physiology, genetics, immunology, and other scientific sub-disciplines have helped to define and broaden the scope of contemporary medicine from the beginning of a research tradition in the modern era.

Medical science benefits from biomedical science because it enables doctors to comprehend the crucial steps involved in infectious diseases brought on by bacteria, viruses, protozoa, and other microorganisms, the impact of body physiology and biochemistry on maintaining health, and the immune system's tolerance or rejection of transplanted tissues. It provides a framework for developing novel methods of maintaining health as well as for testing someone's blood, urine, or tissue for the presence of disease.

The goal of biomedical science is to identify diseases using various techniques. Early diagnosis can save a person's life in many conditions, including cancer. Over the last decade, technologies have been driving the healthcare industry through various innovations in how we find, prevent, and cure diseases. This shouldn't have happened without the massive growth of AI-driven technologies and digitization of healthcare workflows, as a response to more savage global conditions, as well as the rising demand on accessible and quality medical service. Those medical innovations have pushed the envelope of possibility and increased the well-being of millions. This year is no different. Doctors and researchers on the forefront of medicine and technology are enhancing patient care in a number of ways with technology spearheading the initiatives. Here are some medical innovations: bringing diseases to an end with CRISPR Technology, UAV technology for medical supply distribution, IoT for healthcare, and remote patient monitoring.

Recent ML developments promise to significantly enhance the accuracy of diagnosis and the screening for retinal disorders. Systems created using these techniques have shown expert-level accuracy in the detection of a variety of eye disorders, including glaucoma, age-related macular degeneration (AMD), diabetic retinopathy, and other anomalies related to retinal diseases[1]–[3]. But it's not entirely clear how these models affect clinical settings. Many difficulties have been encountered in the past when ML algorithms have been used in computer-assisted diagnosis settings, including over reliance (repeating model errors) and under reliance (ignoring accurate algorithm predictions) [4], [5]. If the computer assisted diagnosis system can explain its black box AI predictions, some of these problems might be avoided [6]). Explainable AI (XAI) aims at decoding the decision of AI (Deep learning/Machine learning) black box to the extent of human-interpretable level. As such we pose the following research questions: (RQ1). How has explainable AI been applied in the sphere of biomedical science? (RQ2) How can deep learning algorithms to classify Liver Disease from a set of patients' records generate further interpretable justification for its prediction results? (RQ3) Can the justification of explainable AI results for predicting the presence or absence of liver disease be visually presented? Our paper aims to contribute to the ongoing research on explainability in line with the desire for understanding of AI predictions in industry.

The subsequent sections of the article are structured as follows, related works section where we delve into AI's diffusion in biomedical science; followed by the next section which explores explainable AI (XAI); followed by the data and method section where we train neural network models and apply XAI algorithms on the results (revealed in the results section); finally, we conclude on the study and summarize our findings.

2 Related works

Many important problems in biomedical decision making can be expressed as binary classification problems. For example, one may wish to identify infants infected with hepatitis C virus from a sample of infants born to infected mothers [7], screen for prostate cancer using prostate-specific antigen [8], or predict which breast cancer patients will respond to treatment based on genetic characteristics [9].

In order to address the methods, techniques and algorithms used for making decisions in biomedicine, let us take into account the following aspects of medical data processing: missing data imputation, diagnostics (classification and prediction), clustering and personalizing the treatment. A previous study predicted missing data, analyzed the nature of data gaps, and filled these gaps using decision tree-based computation techniques and regression approach [10]. Similar outcomes for associative rules mining in medical data were found by another study [11]. In addition, a study adopted Bayesian networks, ANN, and k-means algorithms to predict cardiac disease [12]. However, Bayesian networks are too sluggish for both online diagnostics and processing the vast amounts of data. Y. Tang created a method for paralleling Bayesian networks in response to this [13]. For multi-parameter, massive, and dynamic medical data flows, Bayesian networks should still be used in conjunction with other machine learning techniques, even in the presence of parallelism. Fuzzy logic-based artificial neural network technology is actively employed to analyze a variety of medical data. Thus, a system of quick medical diagnosis based on auto-associative neuro-fuzzy memory was developed in the works [14]–[17]. To increase the accuracy of the classification problem's results, however, is still of uttermost priority. The use of existing techniques and computational intelligence tools to address such issues is further constrained by the issue of imbalanced input data as well as the tiny samples of data manually collected by medical professionals [18].

The cluster analysis is frequently used to identify outliers. In the medical field, outliers refer to variations from the ideal patient circumstances based on the regional protocol and unique traits. Partitioning techniques are among the simplest clustering algorithms. The K-means algorithm creates k clusters that are spread far apart from one another. The assumption (hypothesis) regarding the number of clusters and the variety of the instances in various clusters is the fundamental sort of problem that the k-means method solves. The results of prior research and theoretical considerations may be used to guide the selection of the k number [19].

The decisions made in the healthcare industry generally involve a variety of criteria, many options, flawed data, and varying stakeholder preferences. However, the systemic assessment and the processing of pertinent information, a process that involves the flow of data between numerous components, frequently present problems for the decision-makers. Because of this, decision-makers' reliance on informal judgments or processes can result in poor choices in these situations [20]. The widespread availability of data has sparked a growing interest in methods for extracting useful facts and information from data and decision-making that is data-driven. As a result, the data science field seeks to learn from data and frequently impact decisions to make them increasingly dependable. The Decision Support System (DSS) is a flexible framework used in the artificial intelligence (AI) industry for managing the formalization of human problem-solving and contemplation techniques.

DSS can support the problem-solving process based on two principles, including knowledge and the capacity for reasoning. Overall, the consideration of AI is based on a variety of justifications, including an input and operational point of view, an output and behavioral viewpoint, an evaluation of its relevance, i.e., its ideal performance, and a comparison of its consistency and quality with human performance [21]. In order to represent the framework under consideration, distinct AI methodologies lead to different approaches, for instance, for the management of complex problems, such as the significantly complicated decision-making in the healthcare industry. Another important aspect that was emphasized is the idea of distributing processing power and intelligence among network systems. According to Urdea et al., combining patient statistical data with test results data generated at the point of care can result in a complete dataset that can be effectively used to concentrate fine-grained observation data about a variety of diseases using data analysis at both the individual and population levels [22]. According to research, demographic databases combined with test results might be used to obtain a single dimension, which is equivalent to the population's overall health [23]. The large healthcare data may also be retrieved and applied in prediction-based tasks, which is of extreme significance to decision-making in healthcare. This is done by integrating the aforementioned datasets with mobility patterns, location data, and trends in disease pervasiveness.

Table 1. ML and DL applications

Detection	**Prediction**	**Generation**
Image interpretation	Classification	Design
Text & Speech	Analysis	Visual Art
Abuse and Fraud	Recommendations	Text
Human behavior & Identity	Collective behavior	Music

In recent times, deep learning (DL) has been one of the fast-growing ML fields. It attempts to model abstraction from large-scale data by employing multi-layered deep neural networks (DNNs), thus making sense of data such as images, sounds, and texts. Deep Learning helps provide intelligent answers to complex issues. The structure and operation of the human brain serve as its foundation. Artificial neural networks are used by deep learning to analyze data and make predictions. It has applications in practically every business industry.

Deep Learning is used in a large number of applications that are used on a daily basis, such as the Google translator; in virtual assistants such as Yandex Alice, Apple’s Siri, Microsoft’s Cortana and Google Assistant, which use Deep Learning algorithms for voice recognition; classification of emails and even for security systems that make use of facial recognition. Another of the areas where Deep Learning is applied, is in something as complex as autonomous cars, which every day are closer to becoming a reality.

In the case of factories, for example, it can be used to recognize new parts that have not been previously introduced into the system, since the Deep Learning algorithm has studied other previous photos in which it has been indicated what it is a piece and when a new part has been introduced into the system, it has been recognized as such without having to indicate it.

Another very important application in factories is the intelligent recognition of defects. Once the system has been trained with different defects (shape, size, geometry, etc.), it is possible that the system could recognize new defects because it has learned what it is. It is a very interesting application because of the variability of defects it is common not to be able to categorize all at first.

A flood of biological and medical data, including information about medical imaging, biological sequences, and protein structures, has been amassed in recent decades as a result of advancements in high-throughput technology. This section reviews some effective deep learning applications in the biomedical domains.

- ***Medical image classification and segmentation***

Machine learning has long been a potent tool in the diagnosis or assessment of diseases using medical images. Traditionally, classification (identification of diseases or abnormalities) and segmentation of regions of interest (tissues and organs) in various medical applications rely on manually created discriminative characteristics. Participation of skilled physicians is required in this. The widespread use of machine learning in the medical image domain has been hampered by the complexity and ambiguity of medical images, limited expertise in medical image interpretation, and the demand of vast amounts of annotated data. A number of computer vision tasks, including object detection, localization, and segmentation in natural images, have been successfully completed using deep learning techniques.

For the qualitative and quantitative assessment of medical imaging, the segmentation of tissues and organs is essential. To accomplish precise brain tumor segmentation, Pereira et al. used data augmentation, tiny convolutional kernels, and a pre-processing stage [24]. In 2013 and 2015, their CNN-based segmentation technique took first place and second place in the Brain Tumor Segmentation (BRATS) Challenge. Magnetic resonance images (MRI) and a two-phase training process was used by a study to. demonstrated brain tumor segmentation approach (fully automatic) which took the 2nd place in BRATS 2013 [25]. By using the INbreast and Digital Database for Screening Mammography (DDSM) datasets, their methodology outperformed SOTA techniques at the time in terms of model accuracy and effectiveness [26], [27]. Additionally, deep learning architecture in medical research have been shown to segment the heart's left ventricle from MR data [28], the pancreas from computed tomography [29], the prostate from MRI [30], the tibial cartilage from magnetic resonance imaging [31], and the hippocampus from MR brain images [32], [33]. Through semantic segmentation (the process of classifying or labeling each pixel of an image in order to distinguish various tissues or organs [34], [35]) based on a deep neural network architecture where organs, skeletal muscles, as well as fat in CT scans are vividly distinguished [36]. Also, accurate segmentation findings were achieved by semantic segmentation of MRIs [37], [38].

- ***Genomic sequencing and gene expression analysis***

Genomic sequencing, which establishes the precise arrangement of nucleotides within a DNA molecule, is increasingly essential for many applications, including fundamental biological study, medical diagnostics, biotechnology, forensic biology, virology, and biological systematics. Deep learning application in genomic sequencing is divided into two fields: learning the functional activity of DNA sequencing and DNA methylation.

Three processes make up the biological process of gene expression: transcription, RNA processing, and translation. An RNA molecule called precursor messenger RNA (pre-mRNA), which is a copy of the DNA in the transcribed gene, is produced as a result of transcription. The pre-mRNA is then altered by RNA processing to create a new RNA molecule termed messenger RNA (mRNA). Reading the three-letter (codes) in the mRNA sequence during translation results in the creation of a protein molecule (an amino acid chain) [39]. The alternative splicing field and the prediction of gene expression are the two directions in which deep learning techniques are utilized in the field of gene expression.

3 Explainable AI (XAI)

The goal of XAI is to improve the human understanding of the output of AI systems. The term was initially used in previous studies to indicate how well their system could account for the actions of AI-controlled characters in simulation games [40]. Since researchers began looking at explanation for expert systems in the middle of the 1970s, the explainability problem has been a challenge. The unstoppable spread of AI/ML across all spheres and its critical influence in decision-making processes, while not being able to deliver comprehensive details regarding the chain of reasoning leading to some decisions, predictions, recommendations or actions made by it, are directly responsible for the resurgence of this research topic. Therefore, new AI strategies that can make decisions comprehensible and explicable are required due to societal, ethical, and legal demands.

Demystification of the black-box models is at the heart of XAI, which also implies responsible AI because it can aid in the creation of transparent models. This should take place without affecting the accuracy of the AI models; as a result, accuracy and interpretability must frequently be traded off in AI in general and in ML in particular. Accuracy is intimately related to the quality and amount of the training data, which naturally draws a connection to the data science discipline.

Explainability plays a fundamental role in the justification of AI-based predictions or classifications. It aids in prediction verification, model modification, and for unveiling insights into the problem at hand, thereby leading to more dependable AI systems. The need for explaining AI systems is purported to stem from four (4) reasons. In spite of the fact that the four (4) reasons may appear to overlap, it is believed to capture the core motivations of model explainability. These include: Explaining to Justify (the reason for the specific outcome(s)); Explaining to Control (gain insight into vulnerabilities or defects - debugging); Explaining to Improve (a comprehensible model makes improvement possible by focusing on desired constructs); and Explaining to Discover (revealing the unforeseen) [41].

As purported by research the goals of XAI have been summarized into the concepts evident in figure 1. Literature clearly distinguishes between models that can be understood using external XAI approaches and those that are interpretable by design. This distinction between transparent models and post-hoc explainability is more widely understood than the distinction between interpretable models and model interpretability methodologies. This same dichotomy can be seen in the paper discussed in a previous study, where the authors contrast the approaches used to address the transparent box design problem with those used to address the black-box problem's explanation [43].

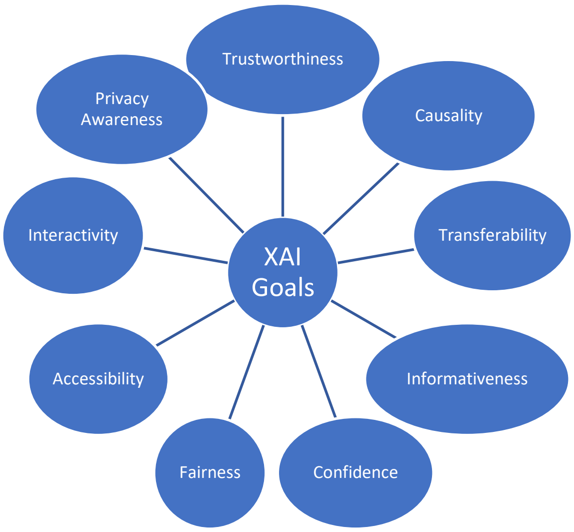


Figure 1: XAI goals [42]

By using a variety of techniques to improve their interpretability, such as text explanations, visual explanations, local explanations, explanations by example, explanations by simplification, and explanations based on feature relevance, post-hoc explainability aims to target models that are not easily interpretable by design techniques.

Here are some XAI methods that have been applied in some real-world tasks, such as autonomous driving and healthcare. These methods develop explainable algorithms to interpret results and improve their decisions or actions according to the task. Recent self-driving systems have adopted interpretation techniques to improve the actions of the autonomous driving system and reduce the risk of a crash. This is also important to increase the trust between humans and AI machines.

- **Explainable decisions for autonomous cars**

In [44], the authors suggested a novel, comprehensible self-driving system that was motivated by human drivers' responses and choices. The suggested solution uses a CNN to extract features from the input image, and a global module to create the scene context and offer information on where the items are in relation to each other. To create the actions and explanations, a local branch is used to pick the scene's most crucial elements and link them to the scene's context. Finally, explanations in visual form are created for the input image. Similar to [45], the authors suggested an architecture for autonomous driving that is aided and trained by humans.

In order to separate the objects from the incoming video stream, the system uses a visual encoder. A vehicle controller is trained to speak commands, such as stopping the automobile when the traffic light turns red, verbally. The controller also creates attention maps to emphasize the key areas and justify their choices. An observation generator is used to aggregate video frames and provide general observations, which must be taken into account while driving, further enhancing the system's robustness. The vehicle controller also receives these observations to help it make better decisions.

- **Explainable medical systems**

AI-based systems have also been used in medical settings in the fields such as drug development and medical imaging, thus produced notable breakthroughs. To help medical professionals by offering helpful explanations so that any expert may grasp a system's predictions, researchers have recently concentrated on explainable medical systems. The authors of concentrated on coronavirus detection from x-ray images [46]. To extract information from the images and determine whether a patient has pneumonia or coronavirus, researchers suggested using a deep convolutional neural network. The infected areas from the x-ray are then highlighted and visual explanations are provided through Grad-CAM [44].

4 Data and method

In this paper, we made use of the Indian Patient Liver Dataset (IPLD) collected from Andhra Pradesh region, a widely known dataset within the ML research community, which comprises observations with 416 liver patient records and 167 non liver patient records [47]. As highlighted in figure 2, the dataset was pre-processed, dropping four (4) unavailable observations, as well as normalization (Min-Max Scalar). The deep learning model is built on Keras-Tensorflow and due to the imbalance in the target values, we integrated Tabular GANs (Generative Adversarial Networks) as a means of oversampling the dataset due to the small sample size.

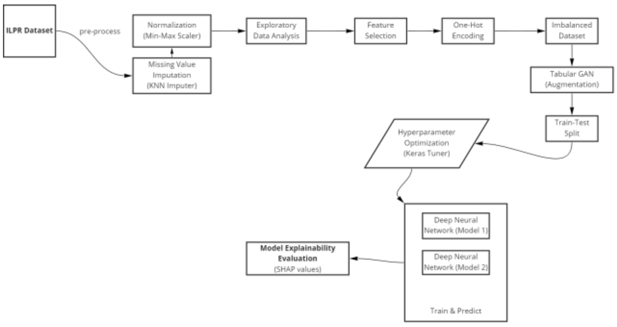

ILPR Dataset
pre-process
Normalization (Min-Max Scaler)
Missing Value Imputation (KNN Imputer)
Exploratory Data Analysis
Feature Selection
One-Hot Encoding
Imbalanced Dataset
Tabular GAN (Augmentation)
Train-Test Split
Hyperparameter Optimization (Keras Tuner)
Deep Neural Network (Model 1)
Deep Neural Network (Model 2)
Train & Predict
Model Explainability Evaluation (SHAP values)

Figure 2: Research workflow

For model interpretability purposes, our study incorporated the SHAP package [48] which was developed as an offspring of research from the University of Washington, and Microsoft Research. Model interpretability is extremely important in AI and it produces end-user trust, delivers insight as to how a model may be improved, as well as supports understanding of the process being modelled [48]. Our study integrated the XAI concept of Shapley Values to shed light on the predictive results obtained by the liver disease detection model. The concept of Shapley Values hails from cooperative game theory and was introduced in 1953 [49]. It is defined as the sum (weighted) of the agents' marginal contributions to coalitions [50]. The three theoretical properties of Shapley values are local accuracy, missingness, and consistency [48]. Marginal contribution is a central component to understanding Shapley values and is defined as the amount by which the evaluation of a submodel increases when a given feature is introduced to the submodel [49]. To formally represent Shapley values as marginal contribution, the formula below is indicated:

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

Where ϕi denotes the average marginal contribution for a player i; N denotes the number of players; v is the game; and S denotes the sets of different coalitions [51]. A Shapley value is representative of a unique quantity that is capable of constructing an explanatory model that locally linearly approximates the original model, given a specific input [52]. From an ML perspective, some studies have adopted Shapley values as a feature selection tool due to its appealing nature with regards to highlighting which features contribute to an obtained output, but in their study, Fryer et al. noted that, in general, the axioms (Efficiency, Null Player, Symmetry, Additivity, and Balanced Contributions) do not provide any guarantee that Shapley values are suitable for feature selection, and may most likely in some cases imply the opposite [49]. They also highlighted that the favorability of Shapley value axioms depends non-trivially on how the Shapley value is appropriated within a particular XAI application.

Shapley values, when applied within a human-centric ML perspective, are capable of shifting the perspective and obtaining insights into client behaviour as well as desires, thereby creating relevant persona profiles which leads towards the trajectory of prescriptive analytics [53]. Shapley values have been applied by previous studies to interpret log anomaly detection systems; to understand client creditworthiness prediction; understand the propensity of clients to buy an insurance policy as well as the risk of churn with respect to an existing customer [52]–[54]. The next section discusses the results of our study.

5 Results

As a means of explaining model predictions, our study utilised SHapley Additive exPlanations (SHAP) and visualises interpretations as SHAP summary plots and SHAP dependence plots. SHAP approximation techniques that exist include Kernel, Deep, and Tree SHAP which are used for kernel-based, deep neural network based, and tree-based models respectively. In order to establish the relationship between features and target variable, initial results from the exploratory data analysis are highlighted in figure 3, via a correlation matrix plot. The following strong positive correlations were established: (1) "Direct_Bilirubin" and "Total_Bilirubin"; (2) "Aspartate_Aminotransferase" and "Alamine_Aminotransferase"; (3) "Albumin" and

"Total_Proteins"; (4) "Albumin and Globulin Ratio" and "Albumin". A negative correlation between the target variable and three features (1) "Total Proteins", (2) "Albumin", and (3) "Albumin and Globulin Ratio", while having a weak positive correlation with the other 8 features.

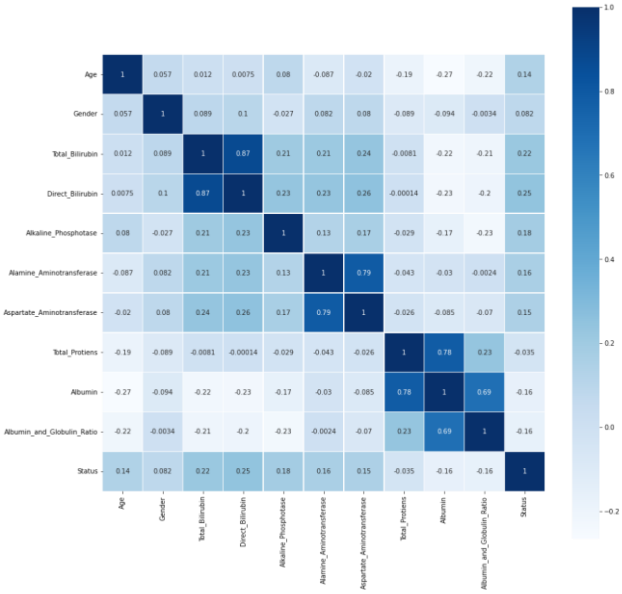


Figure 3: Correlation plot of features and output

Figures 4 and 5 highlight the deep neural network architectures built on Keras-Tensorflow. In order to obtain the best possible training hyperparameters (learning rate, dropout rate, bias vector, neurons, and activation functions), we utilised the RandomSearch feature of the Keras Tuner package (5 and 10 max trials respectively for Models 1 and 2). Figure 3 illustrates the DNN architecture of model 1 and figure 4 highlights that of the hyperparameter tuned model (model 2). The parameter spaces for the hyperparameter tuning process were as follows:

a. Number of Layers – 4

b. Number of Units (Neurons) – value domain = [16 – 512]; step = 16

d. Activation Function – value domain = [ReLU, tanh]; choice step

e. Learning Rate – value domain = [1e-2, 1e-3, 1e-4] – choice step

The binary cross entropy loss as well as the mean absolute error and accuracy metrics were utilized within the Adam optimizer.

Table 3. Sampled data for XAI analysis

Age Category	Gender	Age	Status
Young	F	26	1
	M	18	
	F	29	0
	M	25	
Old	F	58	1
	M	51	
	F	48	0
	M	64	

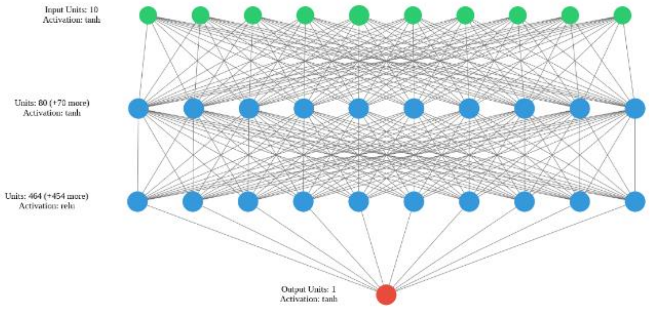


Figure 4: Model 1 - deep neural network architecture

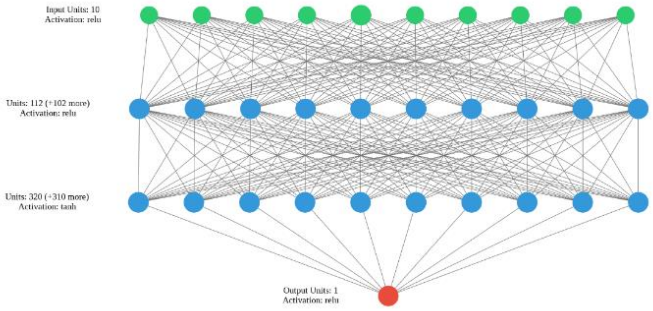


Figure 5: Deep neural network architecture (hyperparameter tuned)

Table 2. 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

Table 2 highlights the model evaluation results of the deep learning models for classifying liver disease patients. Based on Accuracy Model 2 had a slightly higher accuracy but lower precision, recall, and f-measure.

The utilized metrics are calculated as follows (where TP denotes True Positive; TN = True Negative; FP = False Positive; FN = False Negative:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + \ FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F - Measure = \frac{2 \times Precision \ \times Recall}{Precision + Recall}$$

Table 3 highlights the sampled (using purposive sampling), we selected four (4) of the youngest males (2) and female (2) patients, as well as four (4) of the oldest males (2) and females (2) – with one (1) of each sex being an individual with liver disease and the others with no liver disease. The aim of this was to describe the application of SHAP to deep learning models and inferring from the results based on observations within the dataset. In summary, why the model predicted what it predicted.

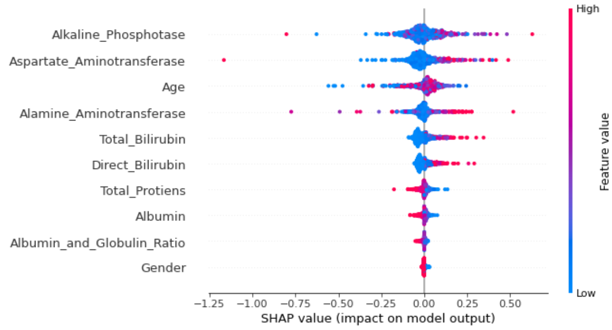


Figure 6: SHAP plots - deep neural network model (best model) - beeswarm plot

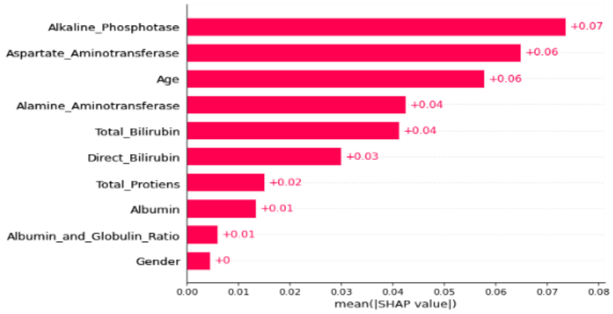
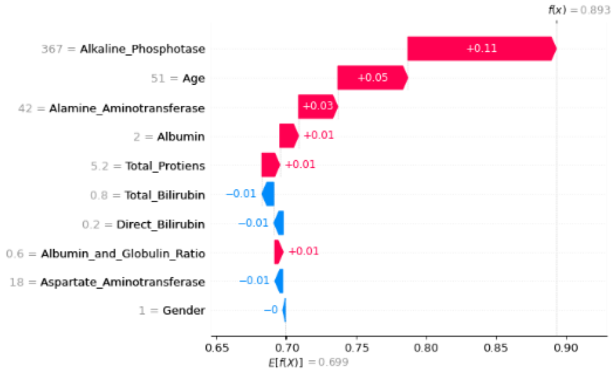


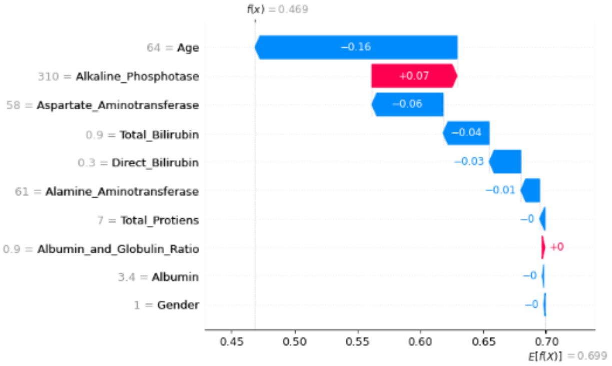
Figure 7: SHAP plots - deep neural network model (best model) - bar plot

In our study we utilized the DeepSHAP functionality due to the fact that it is tailor-made for deep learning models just like ours. Figures 6 and 7 highlight a Beeswarm plot and Bar plot respectively indicating the influence of predictors on the best deep learning model.

The results from our selected sample (from table 4) are presented in figures 8, and 9 for old as well as 10, and 11 for young individuals respectively. The results reveal the impact of certain features on the overall prediction for each selected sample observation; red indicative of the positive contribution and blue indicative of none or negative contribution to the overall outcome. Such results can aid medical staff in understanding how each individual patient's body may react to certain dosage of treatment, thus creating space for personalized treatment.

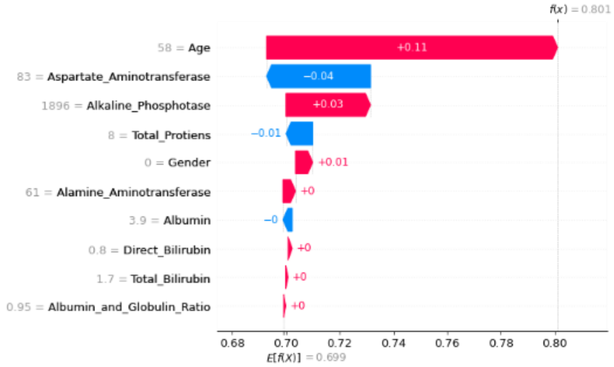


(a) Old male (status = liver disease)

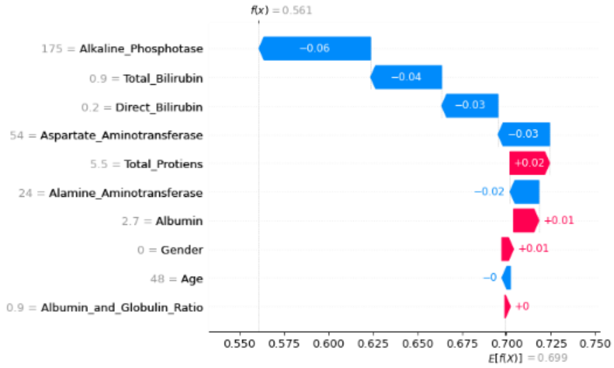


(b) Old Male (Status = No Liver Disease)

Figure 8. Comparative analysis of SHAP plots for two old males (a – with disease and b – no disease)

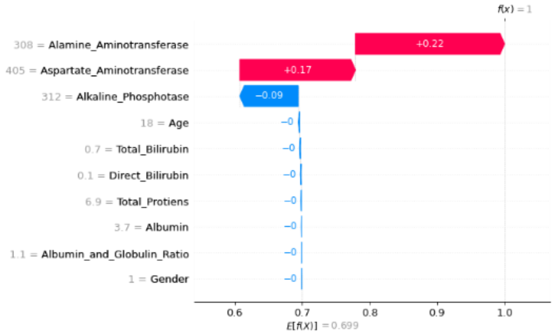


(a) Old female (status = liver disease)

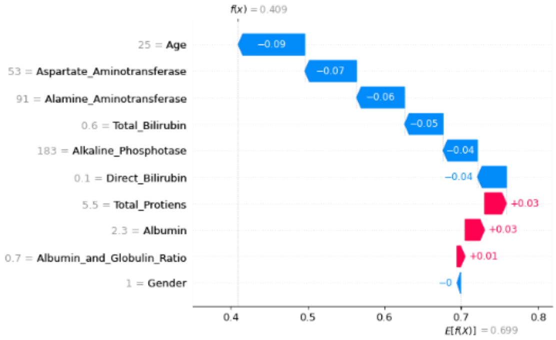


(b) Old Female (Status = No Liver Disease)

Figure 9: Comparative analysis of SHAP plots for two old females (a – with disease and b – no disease)

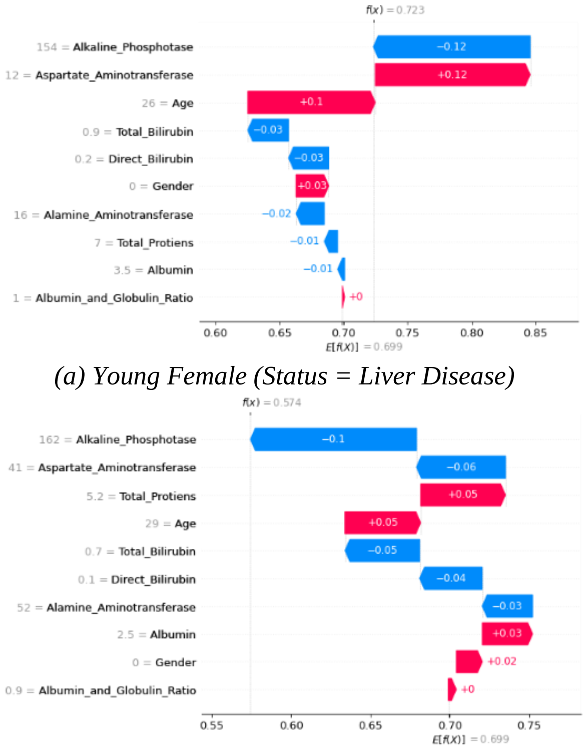


(a) Young male (status = liver disease)
(b)



(b) Young male (status = no liver disease)

Figure 10: Comparative analysis of SHAP plots for two young males (a – with disease and b – no disease)



(a) Young Female (Status = Liver Disease)

(b) Young female (status = no liver disease)

Figure 11: Comparative analysis of SHAP plots for two young females

It can be observed that SHAP plots (from figures 8 and 9, as well as 10 and 11) vary by individual and this gives a more nuanced perspective to model prediction outcomes due to the ability to interpret each predicted outcome and provide personalized solutions to each patient (be it dietary, lifestyle or medical).

In more recent times, with the gradual growth in XAI, there has been some pushback (especially high-stakes decision making) [55]. These conversations will continue as AI research further develops. From our perspective, we conclude that, XAI can be used as a decision support tool provided the model is tested and meets robust real-world and ethical requirements of whichever industry it is needed for. Our research does not claim to propose XAI as the optimal decision support system within healthcare where models play high-stake roles because, in simple terms, XAI is not the remedy for a low performance model within the real world. As such, we recommend end-to-end machine learning which follows current MLOps industry guidelines such as: (a) Efficient Pipelines, Model Re-Training, and Monitoring (Symeonidis et al., 2022); (b) MLOps Maturity Model proposed by John et al. which encompasses Automated Data Collection, Automated Model Deployment, Semi-automated Model Monitoring, Fully-automated Model Monitoring, and as well incorporates governance and security protocols; (c) Responsible AI - Openness to Learning and Changing the Culture, Model Development Preparation, Selection of the Right Tools, Automating the Pipelines, and Monitoring [56], [57]. In summary, the power eXplainable Artificial Intelligence can be experienced, when intrinsically end-to-end AI implementation is done following appropriate MLOps guidelines.

6 Conclusion

As AI continues to gain ubiquity, Explainable AI's relevance is now more than ever essential in all spheres. Primarily in safety-critical domains such as healthcare, the need to interpret AI model predictions will go a long way to support medical treatment as well as personalized medicine.

This study sought to present the applicability of explainability within deep learning models, which have been known as black-box models within the AI sphere. We conducted a research summary on the applications of explainable AI (XAI) in biomedical research and utilized the Indian Liver Patient Dataset as a case study. Furthermore, making use of data-preprocessing, feature selection, data augmentation (with Generative Adversarial Network techniques for Tabular Data), and hyperparameter optimization, we developed deep learning classification models to classify liver disease. In addition, we integrated SHAP (Shapley Values) in interpreting the models, thus establishing model explainability. Finally, we discussed XAI and its implications and made recommendations.

With respect to theoretical implications, our work contributes to the extant literature and conversations on the explainable and interpretable AI paradigm primarily within the healthcare research sphere, i.e. adopting SHAP values. In like manner, our study serves as a contribution to research on data augmentation in the face of inadequate observations for deep learning models. It must be noted that, our research provides practical implications for researchers and health workers to adopt explainable models in supporting decision making process of medical diagnostics and prescription. Practically, our work is relevant to healthcare in deprived areas where trusted AI models (with explainable features) can be deployed on the edge to aid in affordable and mobile healthcare provision.

We recommend future research to reproduce our study within other medical contexts, as well as explore alternative explainable approaches to biomedical healthcare deep learning models. In addition, we recommend future research to delve into developing XAI frameworks or guideliness for healthcare implementation.

Conflict of interests: The authors declare no conflict of interest.

Author contributions: E.A.: Research formulation, Article Writing, and Analysis - Model Explainability Experimentation. D.M.B.: Analysis – ML Model Development, and Article Writing.

References

[1] Grassmann, F., Mengelkamp, J., Brandl, C., Harsch, S., Zimmermann, M. E., Linkohr, B., Peters, A., Heid, I. M., Palm, C., & Weber, B. H. (2018). A deep learning algorithm for prediction of age-related eye disease study severity scale for age-related macular degeneration from color fundus photography. Ophthalmology, 125(9), 1410–1420. https://doi.org/10.1016/j.ophtha.2018.02.037

[2] Nayak, J., Acharya U, R., Bhat, P. S., Shetty, N., Lim, T.-C., & others. (2009). Automated diagnosis of glaucoma using digital fundus images. Journal of Medical Systems, 33(5), 337–346. https://doi.org/10.1007/s10916-008-9195-z

[3] Raman, R., Dasgupta, D., Ramasamy, K., George, R., Mohan, V., & Ting, D. (2021). Using artificial intelligence for diabetic retinopathy screening: Policy implications. Indian Journal of Ophthalmology, 69(11), 2993–2998. https://doi.org/10.4103/ijo.IJO_1420_21

[4] Kohli, A., & Jha, S. (2018). Why CAD failed in mammography. Journal of the American College of Radiology, 15(3), 535–537. https://doi.org/10.1016/j.jacr.2017.12.029

[5] Sayres, R., Taly, A., Rahimy, E., Blumer, K., Coz, D., Hammel, N., Krause, J., Narayanaswamy, A., Rastegar, Z., Wu, D., & others. (2019). Using a deep learning algorithm and integrated gradients explanation to assist grading for diabetic retinopathy. Ophthalmology, 126(4), 552–564. https://doi.org/10.1016/j.ophtha.2018.11.016

[6] Cabitza, F., Rasoini, R., & Gensini, G. F. (2017). Unintended consequences of machine learning in medicine. Jama, 318(6), 517–518. https://doi.org/10.1001/jama.2017.7797

[7] Shebl, F. M., El-Kamary, S. S., Saleh, D. A., Abdel-Hamid, M., Mikhail, N., Allam, A., El-Arabi, H., Elhenawy, I., El-Kafrawy, S., El-Daly, M., & others. (2009). Prospective cohort study of mother-to-infant infection and clearance of hepatitis C in rural Egyptian villages. Journal of Medical Virology, 81(6), 1024–1031. https://doi.org/10.1002/jmv.21480

[8] Etzioni, R., Pepe, M., Longton, G., Hu, C., & Goodman, G. (1999). Incorporating the time dimension in receiver operating characteristic curves: A case study of prostate cancer. Medical Decision Making, 19(3), 242–251. https://doi.org/10.1177/0272989x9901900303

[9] Fan, C., Prat, A., Parker, J. S., Liu, Y., Carey, L. A., Troester, M. A., & Perou, C. M. (2011). Building prognostic models for breast cancer patients using clinical variables and hundreds of gene expression signatures. BMC Medical Genomics, 4(1), 1–15. https://doi.org/10.1186/1755-8794-4-3

[10] Telenyk, S., Czajkowski, K., Bidiuk, P., & Zharikov, E. (2019). Method of assessing the state of monuments based on fuzzy logic. 2019 10th IEEE International Conference on Intelligent Data Acquisition and Advanced Computing Systems: Technology and Applications (IDAACS), 1, 500–506. https://doi.org/10.1109/idaacs.2019.8924315

[11] Dangare, C. S., & Apte, S. S. (2012). Improved study of heart disease prediction system using data mining classification techniques. International Journal of Computer Applications, 47(10), 44–48. https://doi.org/10.5120/7228-0076

[12] Vijiyarani, S., & Sudha, S. (2013). Disease prediction in data mining technique–a survey. International Journal of Computer Applications & Information Technology, 2(1), 17–21.

[13] Tang, Y., Wang, Y., Cooper, K. M., & Li, L. (2014). Towards big data Bayesian network learning-an ensemble learning based approach. 2014 IEEE International Congress on Big Data, 355–357. https://doi.org/10.1109/bigdata.congress.2014.58

[14] Bodyanskiy, Y., Perova, I., Vynokurova, O., & Izonin, I. (2018). Adaptive wavelet diagnostic neuro-fuzzy network for biomedical tasks. 2018 14th International Conference on Advanced Trends in Radioelecrtronics, Telecommunications and Computer Engineering (TCSET), 711–715. https://doi.org/10.1109/tcset.2018.8336299

[15] Perova, I., Brazhnykova, Y., Bodyanskiy, Y., & Mulesa, P. (2018). Neural network for online principal component analysis in medical data mining tasks. 2018 IEEE First International Conference on System Analysis & Intelligent Computing (SAIC), 1–5. https://doi.org/10.1109/tcset.2018.8336299

[16] Perova, I., Litovchenko, O., Bodvanskiy, Y., Brazhnykova, Y., Zavgorodnii, I., & Mulesa, P. (2018). Medical data-stream mining in the area of electromagnetic radiation and low temperature influence on biological objects. 2018 IEEE Second International Conference on Data Stream Mining & Processing (DSMP), 3–6. https://doi.org/10.1109/dsmp.2018.8478577

[17] Perova, I., & Mulesa, P. (2015). Fuzzy spacial extrapolation method using Manhattan metrics for tasks of Medical Data mining. 2015 Xth International Scientific and Technical Conference" Computer Sciences and Information Technologies"(CSIT), 104–106. https://doi.org/10.1109/stc-csit.2015.7325443

[18] Izonin, I., Trostianchyn, A., Duriagina, Z., Tkachenko, R., Tepla, T., & Lotoshynska, N. (2018). The combined use of the wiener polynomial and SVM for material classification task in medical implants production. International Journal of Intelligent Systems and Applications, 10(9), 40–47. https://doi.org/10.5815/ijisa.2018.09.05

[19] Melnykova, N., Shakhovska, N., & Sviridova, T. (2017). The personalized approach in a medical decentralized diagnostic and treatment. 2017 14th

International Conference The Experience of Designing and Application of CAD Systems in Microelectronics (CADSM), 295–297. https://doi.org/10.1109/cadsm.2017.7916139

[20] Baltussen, R., & Niessen, L. (2006). Priority setting of health interventions: The need for multi-criteria decision analysis. Cost Effectiveness and Resource Allocation, 4(1), 1–9. https://doi.org/10.2139/ssrn.943814

[21] Russell, S. J. (2010). Artificial intelligence a modern approach. Pearson Education, Inc. https://doi.org/10.1016/0004-3702(96)00007-0

[22] Urdea, M., Penny, L. A., Olmsted, S. S., Giovanni, M. Y., Kaspar, P., Shepherd, A., Wilson, P., Dahl, C. A., Buchsbaum, S., Moeller, G., & others. (2006). Requirements for high impact diagnostics in the developing world. Nature, 444(1), 73–79. https://doi.org/10.1038/nature05448

[23] Drain, P. K., Hyle, E. P., Noubary, F., Freedberg, K. A., Wilson, D., Bishai, W. R., Rodriguez, W., & Bassett, I. V. (2014). Diagnostic point-of-care tests in resource-limited settings. The Lancet Infectious Diseases, 14(3), 239–249. https://doi.org/10.1016/s1473-3099(13)70250-0

[24] Pereira, S., Pinto, A., Alves, V., & Silva, C. A. (2016). Brain tumor segmentation using convolutional neural networks in MRI images. IEEE Transactions on Medical Imaging, 35(5), 1240–1251.

[25] Havaei, M., Davy, A., Warde-Farley, D., Biard, A., Courville, A., Bengio, Y., Pal, C., Jodoin, P.-M., & Larochelle, H. (2017). Brain tumor segmentation with deep neural networks. Medical Image Analysis, 35, 18–31.

[26] Moreira, I. C., Amaral, I., Domingues, I., Cardoso, A., Cardoso, M. J., & Cardoso, J. S. (2012). Inbreast: Toward a full-field digital mammographic database. Academic Radiology, 19(2), 236–248.

[27] PUB, M. H., Bowyer, K., Kopans, D., Moore, R., & Kegelmeyer, P. (n.d.). The digital database for screening mammography. Proceedings of the Fifth International Workshop on Digital Mammography, 212–218.

[28] Ngo, T. A., Lu, Z., & Carneiro, G. (2017). Combining deep learning and level set for the automated segmentation of the left ventricle of the heart from cardiac cine magnetic resonance. Medical Image Analysis, 35, 159–171. https://doi.org/10.1016/j.media.2016.05.009

[29] Roth, H. R., Farag, A., Lu, L., Turkbey, E. B., & Summers, R. M. (2015). Deep convolutional networks for pancreas segmentation in CT imaging. Medical Imaging 2015: Image Processing, 9413, 378–385. https://doi.org/10.1117/12.2081420

[30] Prasoon, A., Petersen, K., Igel, C., Lauze, F., Dam, E., & Nielsen, M. (2013). Deep feature learning for knee cartilage segmentation using a triplanar convolutional neural network. International Conference on Medical Image Computing and Computer-Assisted Intervention, 246–253. https://doi.org/10.1007/978-3-642-40763-5_31

[31] Liao, S., Gao, Y., Oto, A., & Shen, D. (2013). Representation learning: A unified deep learning framework for automatic prostate MR segmentation. International Conference on Medical Image Computing and Computer-Assisted Intervention, 254–261. https://doi.org/10.1007/978-3-642-40763-5_32

[32] Guo, Y., Wu, G., Commander, L. A., Szary, S., Jewells, V., Lin, W., & Shen, D. (2014). Segmenting hippocampus from infant brains by sparse patch matching with deep-learned features. International Conference on Medical Image Computing and Computer-Assisted Intervention, 308–315. https://doi.org/10.1007/978-3-319-10470-6_39

[33] Kim, M., Wu, G., & Shen, D. (2013). Unsupervised deep learning for hippocampus segmentation in 7.0 Tesla MR images. International Workshop on Machine Learning in Medical Imaging, 1–8. https://doi.org/10.1007/978-3-319-02267-3_1

[34] Schlegl, T., Waldstein, S. M., Vogl, W.-D., Schmidt-Erfurth, U., & Langs, G. (2015). Predicting semantic descriptions from medical images with convolutional neural networks. International Conference on Information Processing in Medical Imaging, 437–448. https://doi.org/10.1007/978-3-319-19992-4_34

[35] Xu, Y., Li, Y., Wang, Y., Liu, M., Fan, Y., Lai, M., Eric, I., & Chang, C. (2017). Gland instance segmentation using deep multichannel neural networks. IEEE Transactions on Biomedical Engineering, 64(12), 2901–2912. https://doi.org/10.1109/tbme.2017.2686418

[36] Lerouge, J., Hérault, R., Chatelain, C., Jardin, F., & Modzelewski, R. (2015). IODA: An input/output deep architecture for image labeling. Pattern Recognition, 48(9), 2847–2858. https://doi.org/10.1016/j.patcog.2015.03.017

[37] Moeskops, P., Viergever, M. A., Mendrik, A. M., De Vries, L. S., Benders, M. J., & Išgum, I. (2016). Automatic segmentation of MR brain images with a convolutional neural network. IEEE Transactions on Medical Imaging, 35(5), 1252–1261. https://doi.org/10.1109/tmi.2016.2548501

[38] Shin, H.-C., Orton, M. R., Collins, D. J., Doran, S. J., & Leach, M. O. (2012). Stacked autoencoders for unsupervised feature learning and multiple organ detection in a pilot study using 4D patient data. IEEE Transactions on Pattern Analysis and Machine Intelligence, 35(8), 1930–1943. https://doi.org/10.1109/tpami.2012.277

[39] Leung, M. K., Delong, A., Alipanahi, B., & Frey, B. J. (2015). Machine learning in genomic medicine: A review of computational problems and data sets. Proceedings of the IEEE, 104(1), 176–197. https://doi.org/10.1109/jproc.2015.2494198

[40] Van Lent, M., Fisher, W., & Mancuso, M. (2004). An explainable artificial intelligence system for small-unit tactical behavior. Proceedings of the National Conference on Artificial Intelligence, 900–907.

[41] Adadi, A., & Berrada, M. (2020). Explainable AI for Healthcare: From Black Box to Interpretable Models.

Scopus. https://doi.org/10.1007/978-981-15-0947-6_31

[42] Nazar, M., Alam, M. M., Yafi, E., & Su'Ud, M. M. (2021). A Systematic Review of Human-Computer Interaction and Explainable Artificial Intelligence in Healthcare with Artificial Intelligence Techniques. IEEE Access. https://doi.org/10.1109/ACCESS.2021.3127881

[43] Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. ACM Computing Surveys (CSUR), 51(5), 1–42. https://doi.org/10.1145/3236009

[44] Xu, Y., Yang, X., Gong, L., Lin, H.-C., Wu, T.-Y., Li, Y., & Vasconcelos, N. (2020). Explainable object-induced action decision for autonomous vehicles. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9523–9532. https://doi.org/10.1109/cvpr42600.2020.00954

[45] Kim, J., Moon, S., Rohrbach, A., Darrell, T., & Canny, J. (2020). Advisable learning for self-driving vehicles by internalizing observation-to-action rules. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 9661–9670. https://doi.org/10.1109/cvpr42600.2020.00968

[46] Brunese, L., Mercaldo, F., Reginelli, A., & Santone, A. (2020). Explainable deep learning for pulmonary disease and coronavirus COVID-19 detection from X-rays. Computer Methods and Programs in Biomedicine, 196, 105608. https://doi.org/10.1016/j.cmpb.2020.105608

[47] Ramana, B., Babu, M., & Venkateswarlu, N. (2012). ILPD (Indian Liver Patient Dataset) Data Set.

[48] Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30.

[49] Fryer, D., Strümke, I., & Nguyen, H. (2021). Shapley values for feature selection: The good, the bad, and the axioms. IEEE Access, 9, 144352–144360. https://doi.org/10.1109/access.2021.3119110

[50] Maniquet, F. (2003). A characterization of the Shapley value in queueing problems. Journal of Economic Theory, 109(1), 90–103. https://doi.org/10.1016/s0022-0531(02)00036-4

[51] Souza, J., & Leung, C. K. (2021). Explainable Artificial Intelligence for Predictive Analytics on Customer Turnover: A User-Friendly Interface for Non-expert Users. In Explainable AI Within the Digital Transformation and Cyber Physical Systems (pp. 47–67). Springer. https://doi.org/10.1007/978-3-030-76409-8_4

[52] Bussmann, N., Giudici, P., Marinelli, D., & Papenbrock, J. (2021). Explainable machine learning in credit risk management. Computational Economics, 57(1), 203–216. https://doi.org/10.1007/s10614-020-10042-0

[53] Gramegna, A., & Giudici, P. (2020). Why to buy insurance? An explainable artificial intelligence approach. Risks, 8(4), 137. https://doi.org/10.3390/risks8040137

[54] Tallón-Ballesteros, A., & Chen, C. (2020). Explainable AI: Using Shapley value to explain complex anomaly detection ML-based systems. Machine Learning and Artificial Intelligence, 332, 152. https://doi.org/10.3233/faia200777

[55] Rudin, C. (2019). Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead. Nature Machine Intelligence, 1(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

[56] John, M. M., Olsson, H. H., & Bosch, J. (2021). Towards mlops: A framework and maturity model. 2021 47th Euromicro Conference on Software Engineering and Advanced Applications (SEAA), 1–8. https://doi.org/10.1109/seaa53835.2021.00050

[57] Matsui, B. M., & Goya, D. H. (2022). MLOps: A Guide to its Adoption in the Context of Responsible AI. 2022 IEEE/ACM 1st International Workshop on Software Engineering for Responsible Artificial Intelligence (SE4RAI), 45–49. https://doi.org/10.1145/3526073.3527591