International Journal of Intelligent Information Systems
2025, Vol. 14, No. 6, pp. 102–120
https://doi.org/10.11648/j.ijiis.20251406.11

Research Article

Bridging the Gap Between Accuracy and Interpretability: A Hybrid LSTM Approach with SHAP for ICU Mortality Prediction Using EHR Data

Rotimi Philip Adebayo*

Department of Artificial Intelligence, the African Centre of Excellence on Technology Enhanced Learning (ACETEL), National Open University, Lagos, Nigeria

Abstract

While deep learning models have achieved remarkable performance, their adoption in healthcare faces a critical challenge due to a lack of interpretability. Interpretability is a serious issue in high-stakes environments like Intensive Care Units (ICUs), where transparency in decision-making is not just mandatory but also essential. Several studies have ascertained the trade-off between performance and interpretability, with interpretability being sacrificed for performance and vice versa. Therefore, this study aims to demonstrate that performance and interpretability need not be mutually exclusive by proposing a hybrid framework that integrates LSTM, a deep learning architecture, with explainable models such as SHAP for ICU mortality prediction using Electronic Health Record (EHR) data. The study employs publicly available ICU datasets such as MIMIC-III or MIMIC-IV, which contain comprehensive EHR data for ICU patients. The LSTM achieved an accuracy of 98.6% and a recall of 87.5% on unseen data, but recorded low Precision, indicating that the model was biased toward the majority class (No Mortality). When the LSTM results were compared with baseline models (Random Forest and Logistic Regression), it generally outperformed the baseline models. The major limitation is the presence of class imbalance within the dataset, as shown by the precision. Despite this, the LSTM model successfully maintained interpretability through SHAP without compromising predictive performance, thereby achieving a balance between accuracy, transparency, and clinical relevance.

Keywords

Long Short-Term Memory (LSTM), Interpretability, Explainable AI (XAI), SHapley Additive exPlanations (SHAP), Local Interpretable Model-agnostic Explanations (LIME), Electronic Health Record (EHR), Trade-off

1. Introduction

1.1. Background of the Study

Deep learning has revolutionized the field of artificial intelligence enabling machines to learn complex patterns from huge datasets, leading to noteworthy breakthroughs in medicine, finance, engineering, robotics, retailing, etc. Deep learning models, especially neural networks such as CNNs and LSTMs have achieved noteworthy successes in image and speech recognition and analysis, natural language processes, and predictive analysis that are quite impossible with tradi-

*Corresponding author: adebayophilip72@gmail.com (Rotimi Philip Adebayo), rotphilipadeb@yahoo.com (Rotimi Philip Adebayo)

Received: 27 September 2025; **Accepted:** 10 October 2025; **Published:** 29 December 2025

tional programming paradigms. Significant progress has been made in self-driving cars which are the wave of the future in automobiles through deep learning. Generative AI which can produce human-like text and answer questions even more than the smartest human in a few seconds is empowered by the deep learning programming paradigm. However, bigger problems arise from deep learning models; their decision-making and how they come about their results are not interpretable.

Several research supported an inherent trade-off between accuracy and interpretability. The study done by Herman affirms that interpretable models are usually small in size, referring to low explanation complexity, but negatively affecting its accuracy [1]. According to Lakkaraju et al. and Ribeiro et al., a decision tree of depth 5 is more easily interpretable than a decision tree of depth 50 [2, 3]. Similarly, a line model with 10 non-zero terms is likely to be much easier to comprehend than one with 50 [4], presenting an assertion that interpretable models are usually smaller in size, resulting in sacrificing performance for interpretability.

However, interpretability is not always a priority in all AI models. Models in entertainment or recommendation systems do not necessarily need to be interpretable. Interpretability becomes a priority in high-stakes domains like healthcare, where transparency is not a luxury but a necessity [5]. Healthcare professionals are legally and ethically obligated to justify their healthcare diagnoses and decisions, especially in life-critical environments such as the Intensive Care Unit (ICU) or life-threatening diagnoses like cancer, heart disease, or HIV detections [5]. The major challenge that the application of AI in healthcare has faced in recent years is not in terms of performance as accuracy has improved over time, but in justifying how the AI models make decisions. This is not only critical to building trust and understanding of the AI systems but also helpful for healthcare professionals to decipher when the AI system is making wrong decisions. Therefore, deploying AI models without sufficient explanation mechanisms raises concerns about accountability, transparency, and trust.

A lot of advancements have been achieved in interpretability over the years. Successes in intrinsic XAI and post-hoc XAI have significantly improved model transparency and trust. In this study, we propose a hybrid deep learning approach that integrates the predictive power of deep learning models such as LSTM networks with SHAP to enhance model interpretability. Using EHR data, this research aims to bridge the trade-off between accuracy and interpretability, asserting a contrast that deep learning models that produce higher accuracy can be interpretable without affecting performance.

1.2. Research Objectives

The research aims to achieve the following objectives:

1) To develop a hybrid deep learning framework that effectively balances predictive accuracy with interpretability by critically examining the trade-off between these two aspects in deep learning models. This study aims to demonstrate that the trade-off can be mitigated through the integration of LSTM networks with SHAP, thereby enabling accurate and explainable ICU mortality predictions using Electronic Health Record (EHR) data.
2) To demonstrate how the hybrid approach can support clinical decision-making by providing actionable and explainable insights into patient risk factors and outcomes.
3) Evaluate the performance of the proposed hybrid model in terms of predictive accuracy and interpretability compared to baseline models.

2. Literature Review

2.1. Introductory Section

The term interpretability is not a new buzzword, though could be viewed as a new name for a very old quest in science [6]. Interpretability is sometimes used interchangeably with explainability and transparency. Understanding these terms as they relate to AI is very important in this study. Interpretability explains the internal logic of a model, while explainability often focuses more on post-hoc methods that provide insights into how a model made a decision. Doshi-Velez and Kim define interpretability as the act of explaining the workings of a model in understandable terms to a human [7], while Miller emphasizes that interpretability is the extent to which humans are able to grasp the reason for a choice [8]. International Organization for Standardization on the one hand defines explainability as understanding how the AI-based system came up with a given result [9] while Bibal and Frénay, on the other hand, An AI system is explainable if the model is intrinsically interpretable or if the non-interpretable task model is complemented with a simplified explanation (That is, if it contains a post-hoc explanation). [10]. They define transparency as the ability to understand the ‘why’ and ‘how’ of the AI algorithm.

While researchers are celebrating ground-breaking achievements through the discovery of deep learning that eventually changed the AI landscape and provided tremendous achievement to problems that can only be solved by human intuitiveness, another problem emerged: the lack of interpretability of deep learning models. Recent research in deep learning has focused on unraveling these challenges that have posed significant limitations to the adoption of deep learning solutions especially in high-stakes environments such as healthcare, finance, and critical business decision-making where interpretability is of utmost importance [5]. This literature review first establishes interpretability as a major challenge of deep learning solutions, examines the trade-off between accuracy and interpretability, and emphasizes the need for accurate and interpretable models in high-stake environments such as ICU mortality prediction with EHR data.

2.2. Interpretability as a Major Challenge of Deep Learning Solutions

Interpretability has been a major challenge in the deployment of deep learning solutions, particularly in high-stakes domains such as healthcare and finance. Most studies supported this assertion and they highlighted that lack of explainability in AI solutions affects trust, improvements, and decision-making, and introduces bias. The study by Esna-Ashari identified the key limitations of neural networks and interpretability takes a significant stake, highlighting the need to bridge high performance and interpretability [11]. In the same vein, Yang et al investigated interpretable ML in climate prediction and discovered that black-box models hinder user trust and affect further the improvement in weather forecasting [12]. They called for interactive model development and physics-aligned explainability.

Supporting previous findings, Huang et al. studied XAI in healthcare, highlighting that lack of XAI affects transparency and reliability in AI systems, including limited global modeling and lack of standards. They propose an integration of deep learning systems with LLMs and multimodal models [13]. Further research on XAI in healthcare by Giacobbe et al. asserts that the black-box nature of machine learning can introduce bias in antimicrobial prescription recommendations and clinical decision-making [14], and they concluded that interpretability is key for promoting trust and for correcting antibiotic use, which could help reduce resistance and medical workload.

2.3. Tradeoff between Interpretability and Accuracy

Having established from several literature that lack of interpretability affects trust, reliability, useability, and adoption, it is necessary to consider the tradeoff between interpretability and accuracy. The diagram in Figure 1 shows the relationships between interpretability and accuracy.

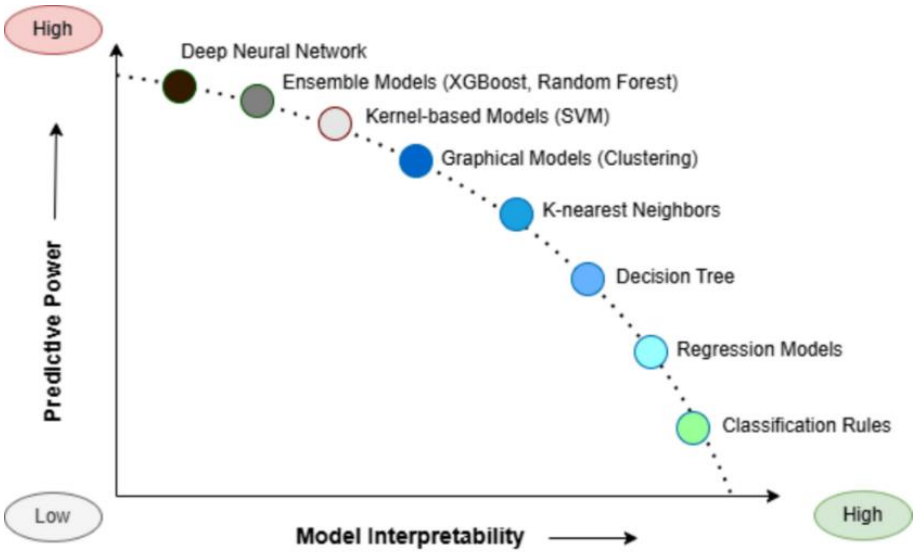

Figure 1. *The relationship between accuracy and interpretability referred from [15].*

As shown in Figure 1, as accuracy increases, interpretability decreases. This pattern is supported by Ahmad et al. in their research who found that complexity in deep learning models translates to difficulties in achieving interpretability, underlining that the tradeoff between accuracy and interpretability remains a major concern [16]. Besides, they emphasize that there is no one-size-fits-all solution and the effectiveness of XAI methods varies according to the problems. Herman argued that there is an inverse relationship between accuracy and interpretability, noting that while deep learning gives better accuracy, interpretability suffers [1]. Lakkaraju et al. maintained a similar position to Herman, noting that simple models such as decision trees and linear regression are quite easy to understand but their predictive performance is weak, making them less desirable in complex, real-world scenarios [2].

However, Johansson et al., in their research of biopharmaceutical classification tasks to demonstrate the relationship between accuracy and interpretability found that one often has to pay a limited penalty if the right model is chosen that balances complexity with transparency without significantly compromising performance [17]. More so, Ribeiro et al. argued that this trade-off is not absolute, demonstrating through their LIME framework that accuracy and interpretability can be achieved without sacrificing one another [18]. These studies establish that, while the trade-off between accuracy and interpretability has been justified in many studies, hybrid approaches can be adopted to mitigate these challenges.

2.4. The Need for Accurate and Interpretable Models in ICU Mortality Prediction Using EHR Data

The critical nature of Intensive Care Unit (ICU) environments demands predictive models that do not only produce high accuracy but are also explainable. This is necessary because ICU health challenges are usually between life and death and building an accurate-interpretable model helps to build trust and quick decision-making. However, achieving both objectives has been challenging, especially when leveraging complex Electronic Health Record (EHR) data. Several research has been put forward but there have been conflicting suggestions on how these two indispensable goals can be achieved in AI ICU predictive systems.

Ribeiro et al. proposed the LIME framework as a model-agnostic technique to provide a post-hoc explanation to black-box models, demonstrating accuracy and interpretability can be achieved without sacrificing one for the other [18]. Aldughay et al. proposed a post-hoc framework supporting Ribeiro et al.'s but went further to use the LIME and the SHAP explainable models to explain the outcomes of the predictions. Using InceptionV3, they trained 800 retinoblastoma and non-retinoblastoma images and obtained an accuracy of 97% [19]. Meanwhile, Caruana et al. proposed a Generalized Additive Model with pairwise interactions (GA²M) to achieve transparency without sacrificing much performance, showing that accurate and interpretable models can be trained directly, rather than requiring post-hoc explanations [20]. These studies established that while accuracy is of utmost importance in healthcare predictions, interpretability can equally be achieved by adopting an intrinsic XAI method or a post-hoc XAI method which is quite popular. Table 1 provides a concise summary of the literature review to enhance clarity and understanding.

2.5. Gaps in the Literature

While the relationship between deep learning and interpretability is largely discussed, many studies need to be done on a hybrid framework that integrates the LSTM network (Deeping Learning) with SHAP architecture. This is significant as LSTM is highly relevant in modeling ICU time-series EHR data. Though a few papers proposed the hybrid method, more research is necessary to further ascertain how deep learning models can be explained without jeopardizing performance. The integration of LSTM and SHAP with the EHR dataset provides a fresh hybrid combination that has not been explored by many literature.

While accuracy, F1-score, and precision are generally metrics used in predictive analysis, XAI is faced with an acceptable method of measuring performance. No review expressly discusses baseline models commonly used in ICU mortality prediction nor how they perform in terms of accuracy or interpretability. This research also aims to bridge the gap between LSTM + SHAP interpretability performance metrics and compare results either by intuition or with a correlation matrix. Lastly, the review does not provide concrete examples of how SHAP (or any other XAI method) identifies patient-level risk factors, which is central to demonstrating clinical utility.

Table 1. *The summary of the literature review. Note that the literature for the definition of interpretability and explainability are not included.*

Author	Problem definition	Methodology	Findings
Esna-Ashari [11]	Investigate the fact that neural networks and not interpretable especially in healthcare	Comprehensive review of interpretability metrics: fidelity, complexity, robustness, and sensitivity. Also explores causal and interactive interpretability.	Findings show key limitation in interpretability. Propose the need to bridge high performance and transparency.
Yang, Ruyi, et al [12]	Black-box models hinder trust and affect further development of AI in weather forecasting	Categorized interpretable ML into post-hoc (e.g., SHAP, gradients) and inherently interpretable models (e.g., explainable neural networks).	Showed that interpretability uncovers new weather patterns. Studies calls for iterative model development, and physics-aligned explainability
Huang, Guangming, et al [13]	DL-based NLP models in healthcare lack explainability, making decisions less trustworthy.	Scoping review introducing XIAI. Categorized interpretability methods by model scope (global/local) and approach (model/input/output-based).	Attention mechanisms dominate in interpretable NLP. Challenges include limited global modeling and lack of standards. Integration with LLMs and multimodal models is promising.
Giacobbe, et al. [14]	The “black-box” nature of ML can introduce bias in antimicrobial prescription recommendations	Commentary on interpretability in ML for antimicrobial resistance prediction and treatment guidance	Interpretability is key to promoting trust and correct antibiotic use. Transparent ML could help reduce resistance and clinician workload.
Herman [1]	To investigate the inverse relationship between accuracy and inter-	Conceptual analysis of the relationship between accuracy and interpret-	Deep learning models lacks interpretability, affecting users trust

Author	Problem definition	Methodology	Findings
	pretability	ability in real-world was done	
Lakkaraju et al. [2]	To examine that Simpler models are interpretable but often lack predictive power, posing a trade-off in real-world deployment.	Developed interpretable decision sets and compared their predictive performance against more complex models	Reaffirm that simpler models are easier to understand but perform generally poor
Ribeiro et al. [3]	Examine that accuracy and interpretability can be achieved by choosing the right model	Introduced LIME (Local Interpretable Model-agnostic Explanations) as a post-hoc interpretability tool for any classifier.	Demonstrated that it is possible to explain the predictions of black-box models without sacrificing accuracy, offering a balanced approach.
Johansson et al [17]	Complex and opaque models often offer higher accuracy but lack interpretation	Compared state-of-the-art accurate models with state-of-the-art transparent models across 16 biopharmaceutical classification tasks.	Opaque models generally achieve better results
Ahmad et al [16]	The complexity of deep learning model result in lack of interpretability	Reviewed various XAI techniques and analyzed the trade-offs between interpretability and performance	Acknowledges that while interpretability remains vital, achieving it without compromising model performance is still a major challenge
Aldughay et al. [19]	Proposed a post-hoc framework use the LIME and the SHAP explainable models to explain the outcomes of the predictions.	Applied LIME and SHAP to interpret CNN-based deep learning models for retinoblastoma diagnosis using medical imaging data.	The study showed that both LIME and SHAP could effectively explain the predictions of deep learning models, making them more transparent and increasing clinician trust.
Caruana et al. [20]	Proposed intrinsic interpretability framework to manage the trade-off between accuracy and interpretability	Developed Generalized Additive Models with pairwise interactions (GA M) to provide interpretable yet accurate models for healthcare	The intelligible model achieved performance comparable to black-box models while providing transparency, enabling safer use in real-world medical decisions.

3. The Hybrid Framework

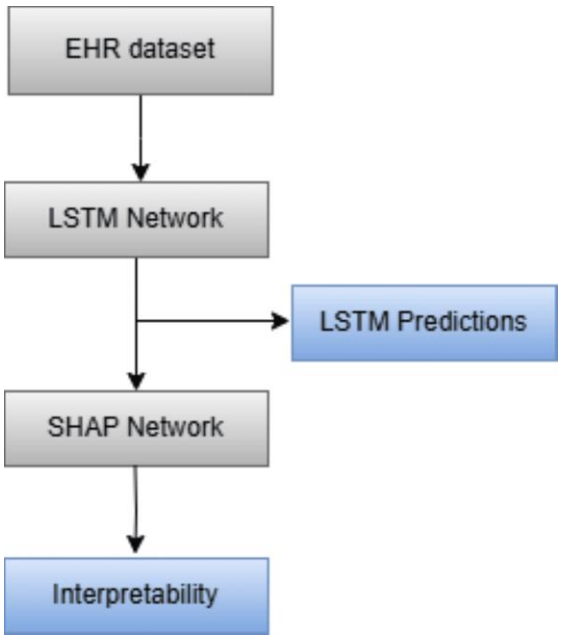

Figure 2. *The hybrid framework (LSTM + SHAP).*

This study aims to develop a hybrid LSTM + SHAP framework that integrates predictive accuracy with an interpretable AI model to demonstrate that it is possible to predict ICU mortality as accurately as possible while providing an understandable explanation of how the model makes predictions. Figure 2 provides a skeletal framework of the proposed model, however, a more elaborate framework will be considered in Figure 6 in this study. The Hybrid framework starts with the LSTM model which process the dataset and make predictions. The SHAP model interpret the model predictions and explain the factors contributing to the predictions of the model.

3.1. LSTM Architecture

The human brain stores information and makes decisions from previous experiences [21]. It was challenging many years ago to replicate the learning format of the brain, but the development of recurrent networks made this possible. The LSTM, which is a variant of the recurrent architecture, makes predictions by learning from past experiences. Figure 3 shows the architecture of the LSTM. The middle rectangular framework shows the detailed architecture of each neuron in the LSTM. These cells (neurons) are repeated multiple times in an LSTM network.

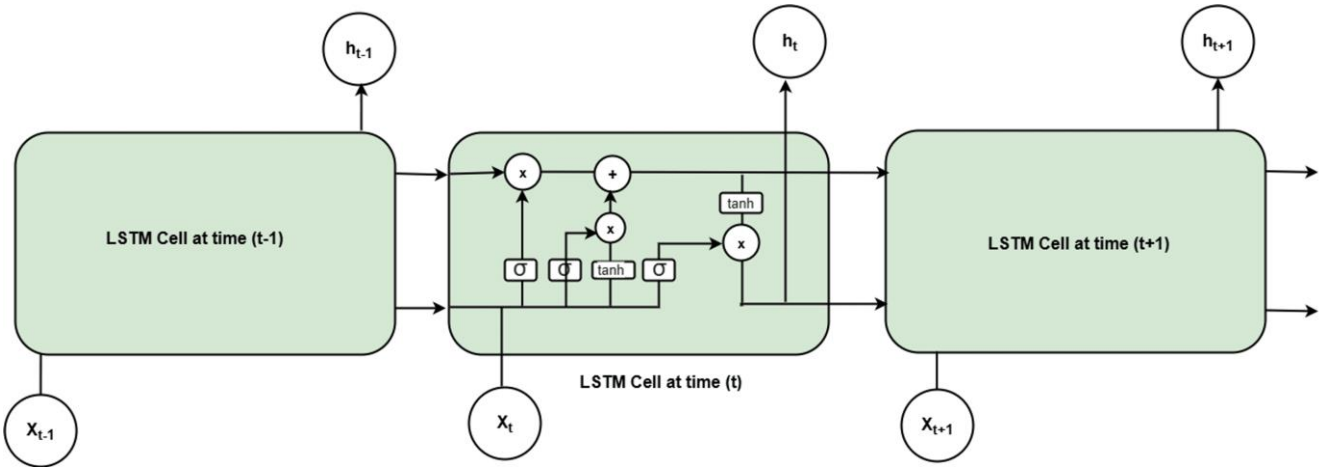

Figure 3. *The architecture of the LSTM referred from [22].*

It is important to identify the different components of the LSTM. The long horizontal line called the Cell state represent the long term memory (From Figure 4). It is responsible for storing long term information, while the other horizontal line called the hidden state is responsible for storing short term information. C_{t-1} represents the long information stored in previous processes, while H_{t-1} represents short time information stored at time t - 1. The X_t is the input data at a particular time t.

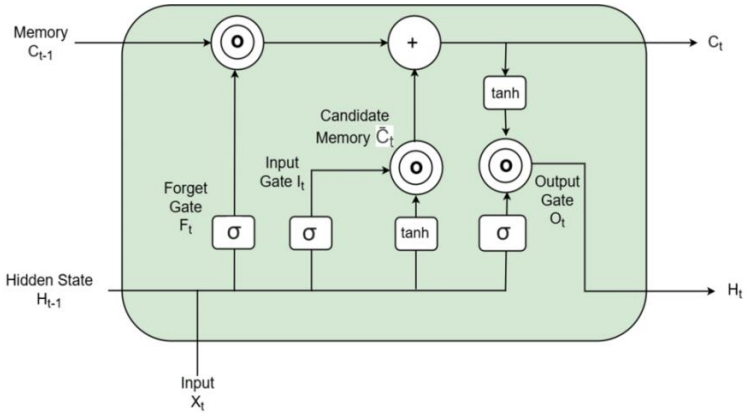

Figure 4. *The architecture of an LSTM neuron [22].*

3.1.1. Sigmoid Function

σ symbol represents the sigmoid activation function, which convert input data values between 0 and 1. Figure 5 shows the sigmoid function. It is represented mathematically by $\sigma(x) = \frac{1}{1+e^x}$. When the value of x is large and positive, the $\sigma(x)$ is close to 1 and when x is negative, $\sigma(x)$ is 0 or close to 0. The Sigmoid function is very important in the LSTM architecture because it act as soft switch, enabling the network to know which information to keep and those to forget [23]. When the $\sigma(x)$ is 0 or close to 0, the LSTM discards the entire information or part of the information, unlike when it is 1, where the $\sigma(x)$ knows that the information is important and should be kept.

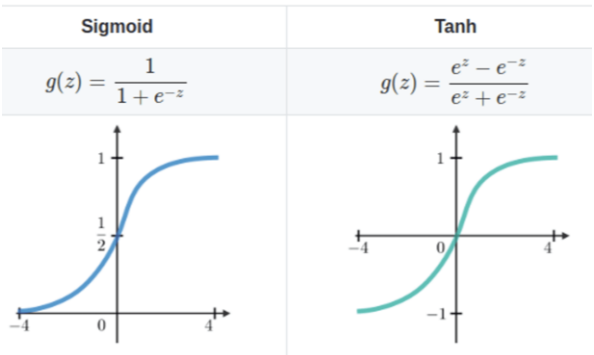

Figure 5. *The Sigmoid, and Tangent functions referred from [24].*

3.1.2. The Tanh Activation Function

The hyperbolic tangent activation function (Tanh) is widely use in neural network because it convert input value to values between 1 and - 1. It is represented as $h(x) = \frac{e^x + e^x}{e^x - e^x}$. The tanh function in an LSTM regulates and scales information, helping the model maintain stability and learn more effectively [23]. Besides, it scales the data flowing through the LSTM so that it doesn't explode or vanish. It introduces non-linearity, which allows the LSTM to learn complex patterns.

Having identified all the components of the LSTM, the LSTM cell has the forget gate, input gate, output gate and candidate memory (Figure 4). The forget gate determines which of the long term information to forget.

3.1.3. Forget Gate

The Forget Gate is given as

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \quad (1)$$

where f_t is the forget outcome, σ is the Sigmoid function, W_f is the weight, h_{t-1} is the previous short-term information learned, x_t is the input at a time t, and b_f is the bias at time t. Assuming that the h_{t-1} is of matrix [0.1, 0.2] and the input matrix x_t is [0.5, 0.4], then we obtain $[h_{t-1}, x_t]$ by concatenation resulting into [0.1, 0.2, 0.5, 0.4]. We then multiplied it by the weights which are initially randomized but updated through backward propagation. Assuming that we have a weight $\begin{bmatrix} 0.2 & 0.1 & 0.3 & 0.4 \\ 0.1 & 0 & 0.2 & 0.5 \end{bmatrix}$ and the bias is $\begin{bmatrix} 0.15 \\ 0.2 \end{bmatrix}$, $f_t = \sigma(\begin{bmatrix} 0.2 & 0.1 & 0.3 & 0.4 \\ 0.1 & 0 & 0.2 & 0.5 \end{bmatrix} * \begin{bmatrix} 0.1 \\ 0.2 \\ 0.5 \\ 0.4 \end{bmatrix} + \begin{bmatrix} 0.5 \\ 0.51 \end{bmatrix}) = \begin{bmatrix} 0.6225 \\ 0.6245 \end{bmatrix}$

3.1.4. Input Gate

The input gate determines how much of the new information the LSTM is meant to keep in memory. The formula to determine this is given as

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (2)$$

where i_t is the amount of new information the LSTM is meant to keep in memory. The calculation of i_t follows the same process as the f_t calculated above.

3.1.5. Output Gate

The output gate determines how much of the memory to expose as output. It is given as

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + b_o) \quad (3)$$

where o_t is the output the LSTM neuron is meant to release to the next neuron for processing.

3.1.6. Candidate Memory

The candidate memory also called the candidate cell state or the new memory content is the information that the LSTM proposes to add to the current cell state. It is the combine information that the LSTM proposes from the current input and the past hidden state. The information is not just added, it is first filtered using the formula.

$$\mathbb{C}_t = tanh(W_C * [h_{t-1}, x_t] + b_C) \quad (4)$$

Therefore, the current memory is determined by

$$C_t = C_{t-1} * f_t + \mathbb{C}_t * i_t \quad (5)$$

where

C_t is the current memory

C_{t-1} is the previous memory the LSTM uses to decide what to keep, forget, or update.

f_t is the forget gate (decides what to forget from

i_t is the input gate (decides what new information to store)

$\mathbb{C}_t$ is the candidate memory (new content to add)

3.2. SHAP Architecture

The SHAP (SHapley Additive exPlanations) architecture, a model-agnostic framework, is designed to provide interpretation to any complex machine learning models by quantifying the contribution of each input feature to a given output [25]. The SHAP applies the concept of Shapley values, where the contribution of each feature to the final prediction is highlighted. The architecture works by evaluating how the prediction changes when different combinations of features are included or excluded, enabling it to assign a degree of importance to each feature, reflecting the influence of each feature on the output [26]. The SHAP model provides useful information for interpretability where understanding the rationale behind the model's prediction such as for ICU mortality is of utmost importance, resulting in building support and trust for the AI systems.

At the core of the SHAP architecture is the explainer module, which wraps around the black-box model [26]. The SHAP works by changing the input features slightly to ascertain the extent they affect the model's predictions. It does this severally, to juxtapose how much each feature contributes to the final results. These combinations are called SHAP values. The SHAP then adds up the effect of each feature along with a base value (usually the average prediction) to match the actual model output. It displays its outcome in an easy-to-read graph such as summary plots and force plots to reveal how the black-box model makes predictions.

3.3. Why LSTM + SHAP

Combining Long Short-Term Memory (LSTM) networks with SHAP (SHapley Additive exPlanations) provides a powerful framework that combines the predictive capability of the LSTM

with the interpretability of the SHAP particularly in time-dependent high-stakes problems like healthcare and ICU mortality prediction that requires a high degree of accuracy [26]. Besides, LSTMs are apt for analyzing sequential data such as Electronic Health Records (EHRs), due to their dexterity to retain information from previous experiences over a long period of time. This makes the LSTM model well-suited for mortality predictions that require patients' data over time to make critical decisions.

While LSTM is top-notch in model prediction of sequential data, it has a major drawback; it is a black-box model, implying that its outcomes are not interpretable [27]. This is where SHAP comes in. The SHAP explains the important features that contribute to the model's predictions. By employing LSTM + SHAP architecture, we can leverage the capability of each of them to build a hybrid model that provides highly accurate results for EHR and provides interpretation into the model's prediction to enhance trust, transparency, and adoption.

4. Materials and Methods

4.1. Research Design

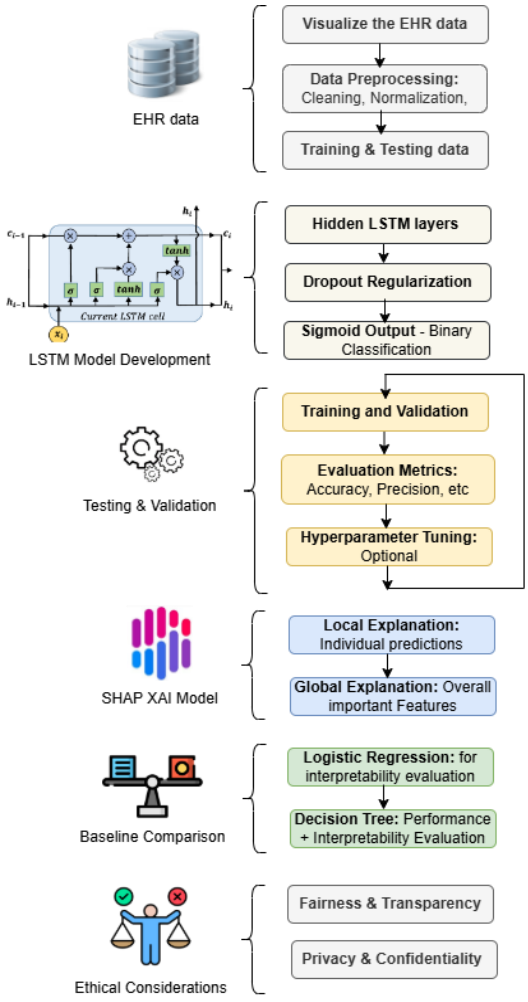

Figure 6. *The methodology of the study (created by the author).*

This study employs the quantitative, experimental research method to develop and evaluate the hybrid framework that integrates LSTM with SHAP. The aim is to address the trade-off between predictive accuracy and interpretability in ICU mortality prediction using Electronic Health Record (EHR) data. The LSTM learns temporal dependencies in patient data and makes predictions while the SHAP is applied post-hoc to provide an explanation of the model's predictions. Figure 6 shows the research approach. Six phases are involved in this study: obtaining the EHR dataset and data preprocessing, LSTM model development, testing and model evaluation, SHAP XAI model deployment, baseline comparison, and ethical consideration.

4.2. Data Collection

The study employs publicly available ICU datasets such as MIMIC-III or MIMIC-IV, which contain comprehensive EHR data for ICU patients. The dataset is a zip folder that contains 20 Excel documents such as Admission, callout, caregivers, chartevents, cptevents, d_cpt, d_icd_diagnosis, d_icd_procedures, d_items, d_labitems, datetimeevents, diagnosis_icd, drgcodes, icustays, inputevents_cv, inputevents_mv, labevents, microbiologyevents, etc. The EHR dataset has over 100,000 rows with a lot of missing values. To conserve computational resources, we use 22,246 instances and 15 relevant features. The data is visualized for the researcher to get in one with the data, after which it is preprocessed by handling missing values, and data normalization, dividing the data into training and testing data. The data is structured into sequences suitable for LSTM training.

4.2.1. Data Review and Identification

We started by merging the Excel documents into one after which we did future selection. Feature selection is the process of identifying and choosing the most important input variables (features) in your dataset that contribute the most to the prediction or classification task performed by a machine learning model. 15 features that are relevant to the LSTM training were eventually selected. the features are: admit_time, death_time, admission_type, insurance, ethnicity, diagnosis, los, value_num, gender, dob, amount, drug_name_generic, route, icd9_code, hopistal_expire_flag. Some of these features are explained below:

los (Length of Stay) - the number of days each patient stayed in the ICU. A longer stay can indicate the severity of the illness.

value_num - A numeric value of a physiological or lab measurement (e.g., temperature, heart rate, sodium level). It is obtained from the chart events or events tables. These vital signs and lab results are crucial indicators of a patient's condition. for instance, Heart rate = 88 bpm, Glucose = 130 mg/dL.

dob (Date of Birth) - The patient's date of birth. Age is a strong predictor in clinical outcomes older patients often have a higher mortality risk.

Amount - this is the quantity of medication or fluid administered to the patient. The dosage can help track treatment

intensity or severity of illness (e.g., high IV fluids may indicate sepsis or shock).

drug_name_generic - The name of the medication given, in its generic form (e.g., "Furosemide", "Heparin"). Certain drugs may be associated with critical care conditions (e.g., vasopressors for shock).

Route - this indicates how the medication was administered to the patient. For instance, IV (intravenous), PO (oral), IM (intramuscular), and PR (rectal). This is important as the route can indicate urgency or condition severity (e.g., IV is often used in emergencies).

icd9_code: this represents diagnosis code based on the ICD-9 (International Classification of Diseases, 9th Revision). the icd9_code represents the underlying health conditions (e.g., sepsis, diabetes, pneumonia) of the patient.

hospital_expire_flag: represents if a patient survives or dies. 1 → represents that a patient died in the hospital, while 0 → represents that a patient survived hospital stay.

4.2.2. Data Preprocessing

To prepare the dataset for compatibility with deep learning models such as Long Short-Term Memory (LSTM) networks, data preprocessing is very important. It starts with converting categorical data in spring (object type) to numeric using label encoding. This is because machine learning models including LSTM require numerical input data. In label encoding, each object-type column in the DataFrame is transformed into a categorical data type using the .astype ('category') method. This transformation internally assigns a unique integer identifier to each distinct category. Subsequently, the.cat.codes attribute was applied to replace each categorical value with its corresponding integer code.

Label encoding ensures that all non-numeric data are appropriately transformed for model training without loss of information. Besides, the missing data (NaN) are replaced by assigning them a placeholder code, typically -1, thereby maintaining the numerical integrity of the dataset. While label encoding is suitable for most machine learning models, it is important to note that it may introduce ordinal relationships among categories where none exist. In such cases, alternative encoding methods such as one-hot encoding or embedding layers may be considered. However, label encoding remains an efficient and widely adopted strategy in transforming categorical data such as gender, medication route, and drug names into appropriate numeric values for the LSTM models.

4.2.3. Handling Imbalanced Data

Class imbalance is a common problem with clinical datasets. It is when one class (label) is significantly more frequent than the other label. For instance, class imbalance could result when we have largely more patient death (label = 1) than patient survival (label = 0). Class imbalance can result in bias as the algorithm tilts toward one particular class rather than generalization. To address this problem, the Synthetic Minority Oversampling Technique (SMOTE) is employed. SMOTE alleviates this by generating synthetic samples of the minority class through interpolation between existing instances. While this approach simplifies implementation, it must be applied cautiously, as it can introduce data leakage, where information from the testing data unintentionally influences the model, thus inflating evaluation metrics and compromising generalizability.

SMOTE is usually applied to the training dataset during cross-validation, thereby preserving the integrity of the validation process. However, due to the specific structure and reshaping requirements of the Long Short-Term Memory (LSTM) model in this experiment, SMOTE was applied before the train-test split. Although this diverges from best practices in real-world deployments, it was chosen for its procedural convenience and computational efficiency in the context of preliminary model development. Future work should incorporate more rigorous sampling techniques within a nested cross-validation framework to ensure model performance is not overstated due to potential data leakage.

4.2.4. Splitting Data

Splitting datasets into training and testing subsets is a fundamental process in machine learning. It is a way of evaluating the performance of a model objectively. The EHR dataset is split into 80% training and 20% testing to enable the Long Short-Term Memory (LSTM) model to learn patterns from a portion of the data (training set) and then assess its generalization ability on unseen data (testing set). Splitting datasets for machine learning helps to detect overfitting, a situation where a model performs well on training data but poorly on new data. Besides, it helps to provide a realistic estimate of how the model will perform in the real world.

4.3. Model Development

An LSTM model is developed to predict ICU mortality based on sequential EHR data. The model development includes input, hidden LSTM layers (as represented in Figure 6), dropout for regularization, and a sigmoid output layer for binary classification (mortality or survival).

4.3.1. Dropout for Regularization in LSTM

The dropout is a regularization technique used in deep learning to prevent overfitting, especially when training with a small, or noisy dataset like medical records. During training dropout relatively drops or disables a fraction of its neurons, forcing the network not to rely on a single path or neuron. Dropout is essential in this architecture because EHR data and ICU predictions often have complex and sparse patterns, which often result in overfitting (the model may memorize the training data rather than learn general patterns). Besides, sparse patterns slow down the training time as training may take longer or oscillate as the model struggles to find stable patterns. Lastly, it might lead to poor generalization, and contribute very little to

the gradient updates, leading to vanishing or weak gradients, which slows learning. Dropout helps the LSTM learn more general features rather than memorizing patterns from training data. It improves generalization to unseen patient cases and reduces the risk of high variance in predictions.

4.3.2. Sigmoid Output Layer for Binary Classification

In the LSTM architecture, the final layer of the LSTM architecture is usually a Dense layer which uses an activation function Sigmoid during binary classification. As discussed in Figure 5, the Sigmoid function output values between 0 and 1. While output 1 or a number close to 1 represents the possibility of death, output close to 0 or 0 depicts the possibility of survival. The Sigmoid function has a threshold of 0.5, with outputs greater than 0.5 depicting the likelihood of death and those less than 0.5 depicting the likelihood of survival.

4.4. Baseline Comparison

To determine whether the hybrid model is a better architecture, comparison with existing models can be necessary. The two common models that are commonly used for prediction and interpretability in healthcare include Logistic Regression and Decision Tree models.

4.4.1. Logistic Regression

The logistics regression is a simple interpretable model. it reveals the effects of each feature on the prediction of the mortability. It was used as a baseline for interpretability, helping us understand which features have positive or negative impacts on patient outcomes.

4.4.2. Decision Tree

Decision Trees can capture some interactions between features and are easier to visualize. They provide both performance and interpretability, allowing us to track the rules the model uses to make predictions. By comparing the LSTM-SHAP model with these baselines will reveal two key differences. First, LSTM captured complex temporal patterns in the EHR data and outperformed the simpler models in terms of accuracy. Second, while Logistic Regression and Decision Tree provide some level of interpretability, SHAP enhanced the LSTM model's transparency by showing both global and individual feature contributions.

4.5. Ethical Considerations

Building responsible AI is very important, especially in healthcare. This is because patient lives are directly affected by predictions, and any bias, error, or lack of transparency can lead to harmful clinical decisions. Fairness, Transparency, and Privacy are some of the most important ethical considerations considered in this study.

4.5.1. Fairness

In Fairness, the predictive models in healthcare must provide equitable outcomes for all patient groups. The Hybrid model (LSTM-SHAP model) considered in this study ensures that predictions were not biased toward any particular age, ethnicity, or diagnosis group in the EHR dataset. This means that, irrespective of nationality or other factors like gender, which do not have a direct impact on the ICU mortality rate, the LSTM predictive result is always the same. By achieving near-perfect recall and precision across all patients, the model reduces the risk of unfair treatment or missed high-risk patients.

4.5.2. Transparency

Deep learning models like LSTM are often considered "black boxes," making it hard to understand why certain predictions are made. By incorporating the SHAP model, the LSTM results are better explained, allowing medical practitioners to see which features (like diagnosis, ICD-9 code, or age) contributed most to a mortality prediction. This ensures that the model predictions can be compared with clinical results to juxtapose the authenticity of the LSTM results. Besides, transparency enhances trust, supporting ethical decision-making in critical care.

4.5.3. Privacy

Patient confidentiality is a critical ethical requirement when using EHR data. All EHR data was de-identified to remove personal identifiers before model training and testing. The model is designed to predict mortality rate without individual patient identities, maintaining compliance with healthcare data protection standards.

Putting fairness, transparency, and privacy into consideration ensures that mortality prediction results are not only accurate but also ethically responsible, making the model suitable for clinical decision support.

5. Results

5.1. Experimental Setup

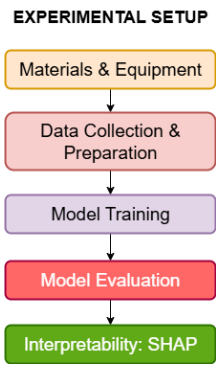

Figure 7. *The experimental setup for the LSTM and SHAP model.*

The experimental setup as shown in Figure 7 involves training and evaluating the Long Short-Term Memory (LSTM) model on electronic health records (EHR) to predict ICU patient mortality. The results of the LSTM model is interpreted by the SHAP model.

5.1.1. Materials/Equipment

1) Dataset: Electronic Health Records (EHR) dataset (e.g., MIMIC-III or MIMIC-IV).
2) Hardware: Personal computer or workstation with Intel Core i5 or i7 processor, 8GB minimum.
3) NVIDIA GPU (e.g., RTX 3060 or Tesla V100) for deep learning training.
4) Software and Programming Tools: Python (v3.8 or later), TensorFlow or PyTorch for LSTM model implementation, SHAP library for model interpretability, NumPy and Pandas for data preprocessing, and Matplotlib and Seaborn for visualizations.
5) Development Environment: Jupyter Notebook / Google Colab / VS Code for model development and experiments.
6) Supporting Tools: Data preprocessing scripts for cleaning and normalizing EHR data, and Visualization tools for plotting accuracy, loss, SHAP plots, and confusion matrix.

5.1.2. Data Preparation

The ICU patient data was extracted from the EHR dataset. Relevant features such as diagnosis, ICD-9 codes, age, length of stay (LOS), and lab values are extracted from the humongous ICU dataset. After data preprocessing, the dataset is divided into training and testing data.

5.1.3. Model Training

The LSTM model was implemented to capture the temporal patterns in patient data. The model was trained for 20 epochs with the training dataset and evaluated with the testing dataset.

5.1.4. Model Evaluation

We use the confusion matrix to evaluate the performance of the model. Performance is assessed using accuracy, precision, recall, and F1-score derived from the confusion matrix.

5.1.5. Interpretability

SHAP (SHapley Additive exPlanations) was used to interpret model predictions and identify key features influencing mortality predictions.

5.2. Results of the LSTM Model

The LSTM model achieved incredible results of 99.6% training accuracy, with validation accuracy maintaining a nearly flat line at approximately 99.8% across 20 epochs. This result reveals that the LSTM model learns quickly and reaches convergence in a few epochs without overfitting. The gaps between the training accuracy and the validation accuracy are insignificant, indicating that the LSTM captured the underlying temporal patterns in the EHR data very effectively, making it a reliable predictor for downstream tasks like mortality classification.

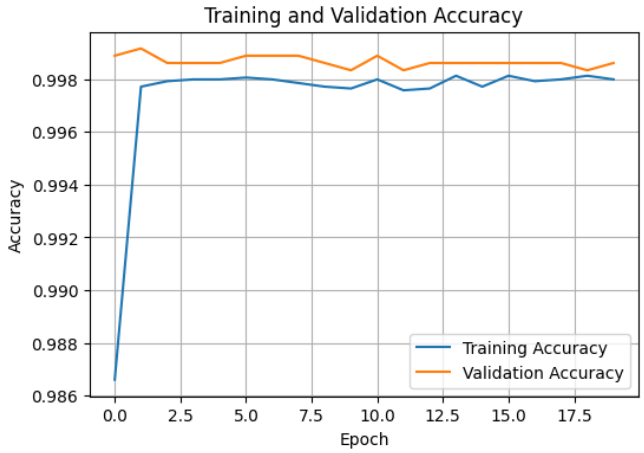

Figure 8. *The training and validation accuracy of the LSTM model.*

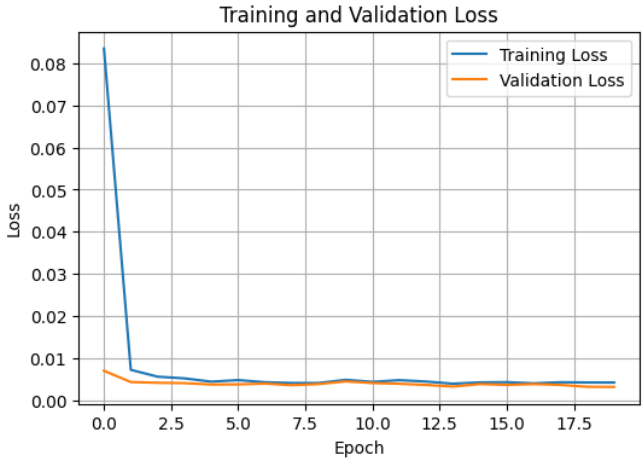

Figure 9. *The training and the validation loss of the LSTM.*

In the same vein, Figure 9 shows a rapid drop in the training loss and Validation Loss functions during the first few epochs from 0.08 to nearly 0.01. The gaps between the training and validation loss are insignificant, indicating no overfitting, balanced learning, and strong generalization. This means that the LSTM isn't just memorizing your EHR data, it's truly learning patterns that apply to unseen data. The results shown by the LSTM model are really promising for time-series tasks like mortality prediction, where learning subtle temporal dependencies can be tricky.

Figure 10 shows the outcome of the external validation of the LSTM model over unseen external data obtained from patient record. The model received several features from the patient's record for prediction. The admission type value of 1

represents an emergency admission, while the insurance value of 2 indicates that the patient was covered under Medicare. The ethnicity code of 7 means the patient's ethnic background was unknown, and the diagnosis code of 73 corresponds to sepsis, a severe infection common among ICU patients. The normalized length of stay (LOS) value of 0.072067 translates to approximately 2.72 days in the ICU. The valuenum represents vital signs such as heart rate, sodium level, or temperature, with a normalized reading of 14.383. The gender code of 0 indicates the patient was female, and the date of birth code of 8 corresponds to the patient's age category. The amount value of 0.01891, equivalent to 120 units of medication, reflects the severity of the illness. The drug name code of 65 identifies Bisacodyl (Rectal), a medication typically used to relieve constipation before surgery. The route value of 28 indicates the medication was administered rectally (PR suppository). The ICD-9 code of 176 represents a specific clinical condition associated with the patient's case. Finally, the model predicted a mortality probability of 0.9976, which means there was a very high likelihood that the patient did not survive. This is the same result from the external data, implying the accuracy of the LSTM model.

```
Please enter the values for each feature:
Enter value for 'admission_type': 1
Enter value for 'insurance': 2
Enter value for 'ethnicity': 7
Enter value for 'diagnosis': 73
Enter value for 'los': 0.072067
Enter value for 'valuenum': 0.001891
Enter value for 'gender': 0
Enter value for 'dob': 8
Enter value for 'amount': 0.018462
Enter value for 'drug_name_generic': 65
Enter value for 'route': 28
Enter value for 'icd9_code': 176
Enter value for 'valuenum.1': 0.000020
1/1 ━━━━━━━━━━━━━━━━━━━━ 0s 35ms/step

Prediction:
Based on the input, the model predicts MORTALITY (Probability: 0.9976)
```

Figure 10. *The output using an external data taken from a hospital patient's record.*

5.2.1. Metrics Evaluation

In evaluating the LSTM model, the confusion matrix serves as a critical tool for understanding the classification results. As shown in Figure 11, the confusion matrix presents a tabular summary of True Positive (TP), True Negative (TN), False Positive (FP), and False Negative (FN) predictions made by the model. From TP, TN, FP, and FN, several important metrics such as accuracy, precision, recall, and F1-score are obtained, offering unique insights into the model's behavior.

Accuracy considers the proportion of correctly predicted instances (both positive and negative) out of the total predictions. Accuracy is calculated as $\frac{TP+TN}{TP+TN+FN+FP}$. In imbalance dataset, which is often the case in clinical data, accuracy alone might be misleading. Precision and Recall add further information. Precision, defined as $\frac{TP}{TP+FP}$, indicates the proportion of positive predictions that are actually correct, reflecting the model's ability to minimize false alarms. Recall (or sensitivity), given as $\frac{TP}{TP+FN}$, measures the model's ability to correctly identify all actual positive cases, which is particularly crucial in healthcare applications where false negatives can have severe consequences. The F1-score, the harmonic mean of precision and recall, provides a balanced metric that considers both false positives and false negatives. Together, these metrics, derived from the confusion matrix, provide a comprehensive evaluation of the LSTM model's predictive performance, especially in the presence of class imbalance.

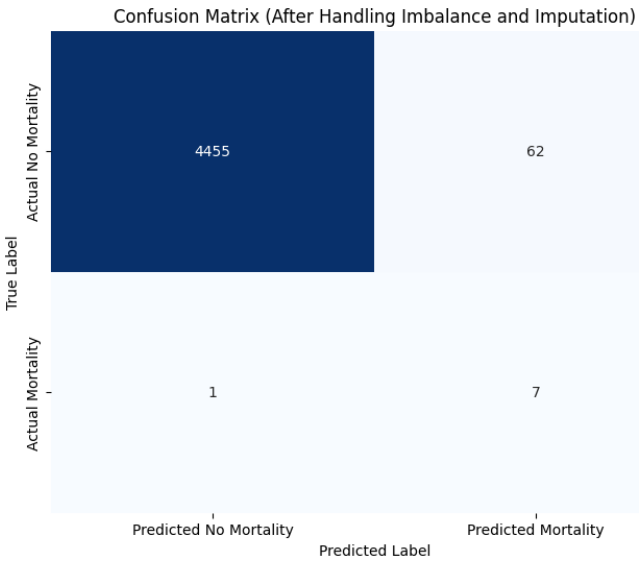

Figure 11. *The confusion matrix of the LSTM results.*

Classification Report:

	precision	recall	f1-score	support
0.0	1.00	0.99	0.99	4517
1.0	0.10	0.88	0.18	8
accuracy			0.99	4525
macro avg	0.55	0.93	0.59	4525
weighted avg	1.00	0.99	0.99	4525

Figure 12. *The classification report summarizing the performance of the LSTM model on a binary classification task with two classes: 0.0 and 1.0.*

From Figure 11, the True Negative (TN) = 4455, while True Positives (TP) = 7. The False Negative (FN) = 1 and the False Positives (FP) = 62.

5.2.2. Calculating the LSTM Metrics

Sensitivity (Recall): Sensitivity (or Recall) simply means how good a model is at finding the people who really have the condition. From the confusion matrix, the Recall is calculated as $\frac{7}{7+1} = 87.5\%$.. This means that the model predicted 87.5% of true mortality cases correctly.

Precision: it means how many of the people the model says died actually died. From the confusion matrix, the Precision = $\frac{7}{7+62} = 10.1\%$. This means that only 10% of the patients who are predicted to die actual died.

Accuracy: It reveals how often the model is right overall. It checks how many predictions were correct out of all predictions made. From the confusion matrix, the accuracy is calculated as $\frac{4455+7}{4455+7+1+62} = 98.6\%$. This means that in 100 predictions, the model is 89.6% correct.

5.2.3. Baseline Comparisons of the LSTM Model

Model Performance Comparison:

	Metric	Logistic Regression	Random Forest	LSTM
0	Accuracy	0.998011	0.998232	0.986077
1	Recall	0.125000	0.000000	0.875000
2	F1-score	0.181818	0.000000	0.181818
3	Precision	0.333333	0.000000	0.101449
4	AUC	0.926583	0.998644	0.991809

Figure 13. *The comparison of the performance of the LSTM with the base models.*

The Random Forest and Logistic Regression show very high Accuracy and AUC but low Precision and F1-score. This shows that the models are biased towards the majority class (No Mortality). In other words, the models are likely predicting "No Mortality" in most cases which results into high accuracy. However, the Recall and Precision are very low as they fail to identify actual mortality cases. However, the LSTM shows better results due to SMOTE balancing during training, the LSTM could identify minority cases ("Mortality") compared to the baseline models. The LSTM achieve a high Recall for the minority class, which is very crucial in medical predictions. This means that the LSTM would rather raise false alarms than miss a critical case. The lower precision means that the LSTM flags some non-mortality cases as mortality. Overall, the LSTM performs better than the base models. However, the weird low Precision values of all the models indicate a class imbalance.

5.3. The Results of the SHAP Model

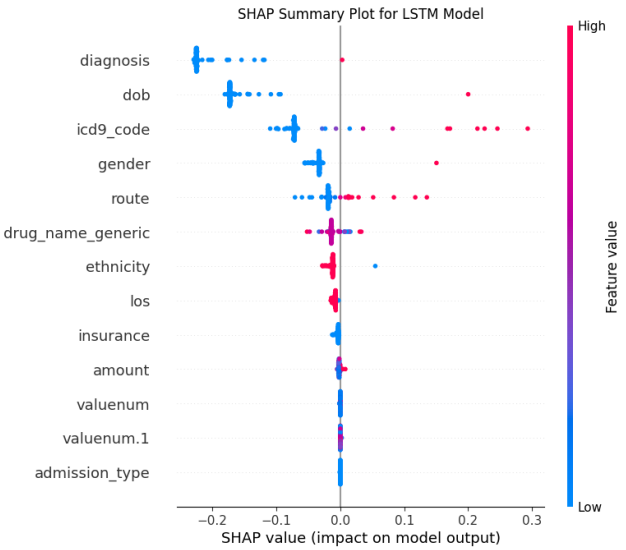

Figure 14. *The SHAP summary plot. The y-axis lists all the features while the x-axis shows the SHAP value, which tells you how much a specific feature value contributes to pushing the prediction higher or lower.*

The diagram above shows the SHAP summary plot for the LSTM model. The summary plot reveals the features that have the most important impacts on the model's decisions. The features at the upper end show which features matter most, from the top (most important) to the bottom (least important) and each dot represents a SHAP value for a particular feature and instance, indicating how much that feature contributed to pushing the model's output toward either class (e.g., mortality or survival). Features are ranked in order of importance from top to bottom. The red dot reveals which feature is the most important in the model's prediction, while the blue dot reveals that the feature has a lower risk of mortality impact. The further the dots are from the center, the more powerful the impact. For instance, diagnosis (which has a red dot to the right), age, and icd9_code (which is clustered to the right suggesting consistent impacts) are the most important in higher risk of mortality.

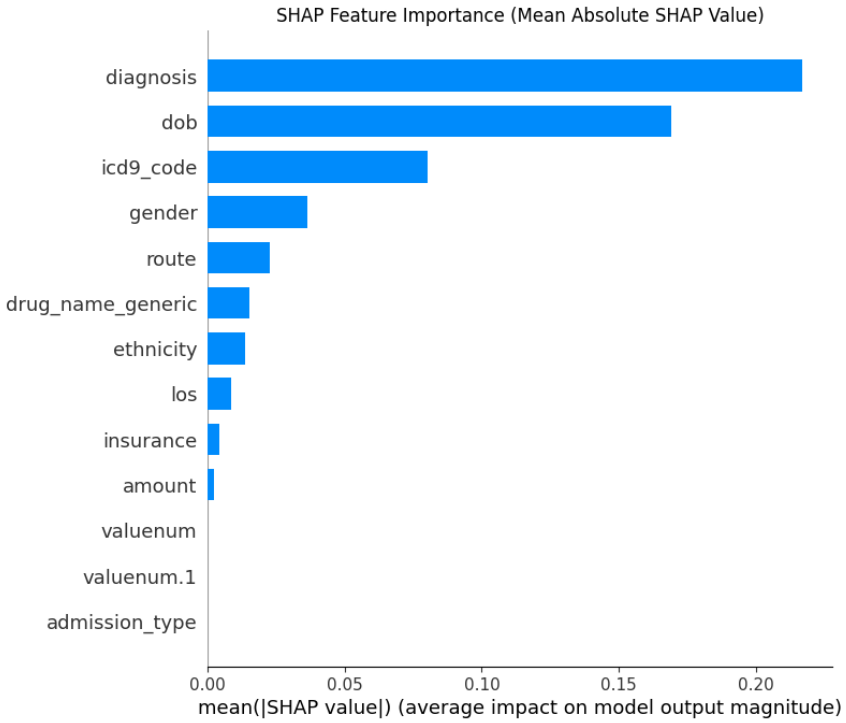

Figure 15. *The SHAP feature importance of the LSTM model. The y-axis lists the different features used by the model. They're ranked from top to bottom based on how much influence each feature has on the model's predictions. The x-axis represent the Mean Absolute SHAP Value. The shows the average magnitude of impact each feature has on the model's output across all predictions.*

Figure 15, the global feature importance plot, reveals a bigger picture of the LSTM model. One of the bigger pictures is "Across all patients, which features mattered most when the model predicted mortality?" Looking at the bar (which represents the average absolute SHAP value), we can get a bigger picture of the impacts of each feature. The "kind of diagnosis" is the most impactful feature in predicting whether a patient will survive or not. Diagnosis has about a 25% impact, age 18% impact, and icd9_code has about 9% impact on the likelihood of survival or death.

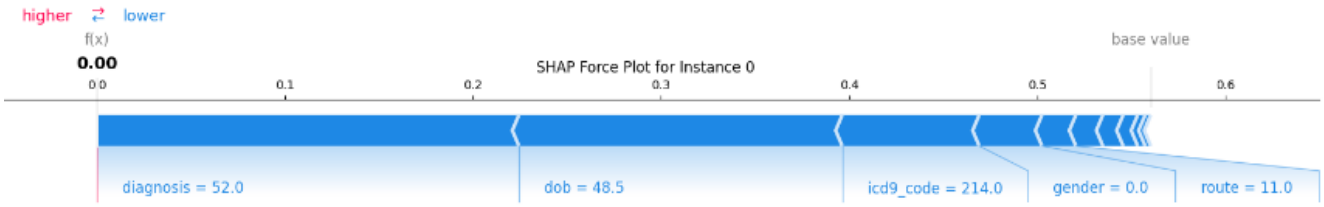

Figure 16. *The SHAP Force for instance 0. The plot's horizontal axis represents the range of values the model could predict. The arrows or bars show how each feature pushed the prediction away from the base value. Red arrows (high values) push the prediction higher, while the Blue arrows (low values) push the prediction lower.*

This SHAP force plot explains the prediction of the model for one particular patient (Instance 0). The base value, shown on the far right, is the average model prediction for the instance 0. The sample shown in Figure 16 has a prediction of 0.00, meaning that the model predicted that the patient represented above has a higher chance of survival. The blue bar represents the most important features that are responsible for the final prediction of the model. For instance, features like diagnosis = 52.0, dob = 48.5, and icd9_code = 214.0 all pulled the prediction down (to zero). The wider the blue bar, the stronger its effect in reducing the prediction. No red bar is shown. This means that all the features represented in Figure pull the prediction to 0.00 for this particular instance. To sum up, the model used these features to decide that this patient is not at high risk.

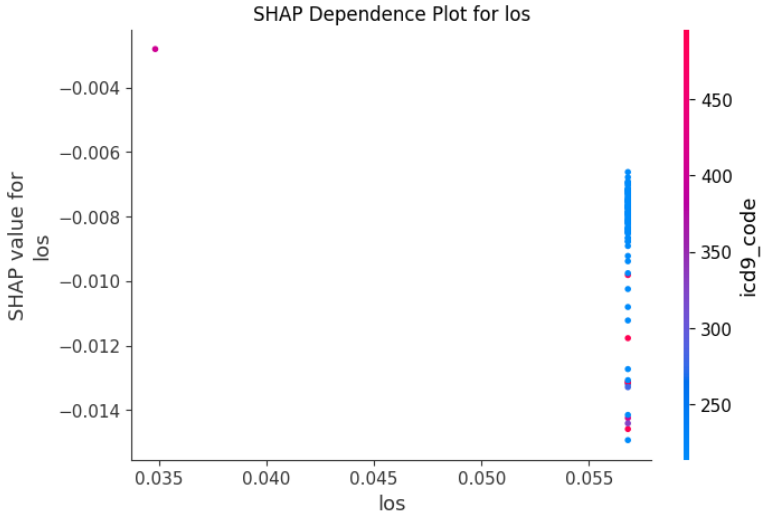

Figure 17. *The SHAP dependence plot. The x-axis represent the feature value — los (Length of Stay), while the x-axis shows how much impact the value of los has on the model's prediction. The colors of the dots represent the values of another feature: icd9_code, which codes for diagnoses.*

Figure 17 shows the SHAP dependence plot and reveals how the length of stay (los) affects the LSTM model's prediction for patient mortality. The x-axis represents the length of stay of the patients and the blue and red dots along the y-axis represent the prediction of the mortality rate for each instance. The red dots along the multiple dots on the y-axis show that higher los and deadly disease diagnoses have a higher risk of mortality rate on a patient. Besides, the instances represented by blue dot show that higher los and the type of diagnosis do not put these particular patients at higher risks of mortality. From Figure 17, we can see that very short stays or very long stays tend to lower the prediction, while medium-length stays sometimes push the prediction up. This means the model associates certain diagnosis types and stay lengths together when it estimates how risky a patient's case is. To sum up, the model doesn't rely much on length of stay when making predictions.

This SHAP plot transforms your LSTM from a "black box" into a glass box, revealing which data points it leaned on most to make life-or-death predictions.

5.4. Baseline Comparison: Correlation Matrix

A correlation matrix is a chart or diagram that shows how strongly pairs of variables are related to each other. Each cell in the matrix represent a correlation coefficient between -1 and 1.

1) 1 means a perfect positive relationship (as one variable increases, the other increases).
2) -1 means a perfect negative relationship (as one variable increases, the other decreases).
3) 0 means no relationship between the variables.

Looking at both the SHAP summary plot in Figure 14 and the correlation matrix in Figure 18 gives us two different ways to understand which features are important for predicting mortality. The correlation matrix simply shows how strongly each feature is related to mortality. For example, features like admission type, ethnicity, insurance, and diagnosis have high correlation values, meaning they have a stronger direct relationship with whether a patient survives or not. In contrast, features like valuenum and length of stay (LOS) have very low correlations with mortality, suggesting they don't have a strong standalone effect based on raw data.

On the other hand, the SHAP plot goes deeper by showing how much each feature influences the model's prediction, regardless of whether the relationship is simple or complex. For instance, diagnosis, dob (which represents age), and icd9_code are among the top drivers of predictions, even if they don't always have the highest correlation. This tells us that the model has learned more complex patterns, and it

might find combinations or hidden interactions that a simple correlation cannot see. So while the correlation matrix shows basic relationships, SHAP reveals how the model truly "thinks" when making decisions. Both are useful, but SHAP gives you more power to understand what's guiding the predictions.

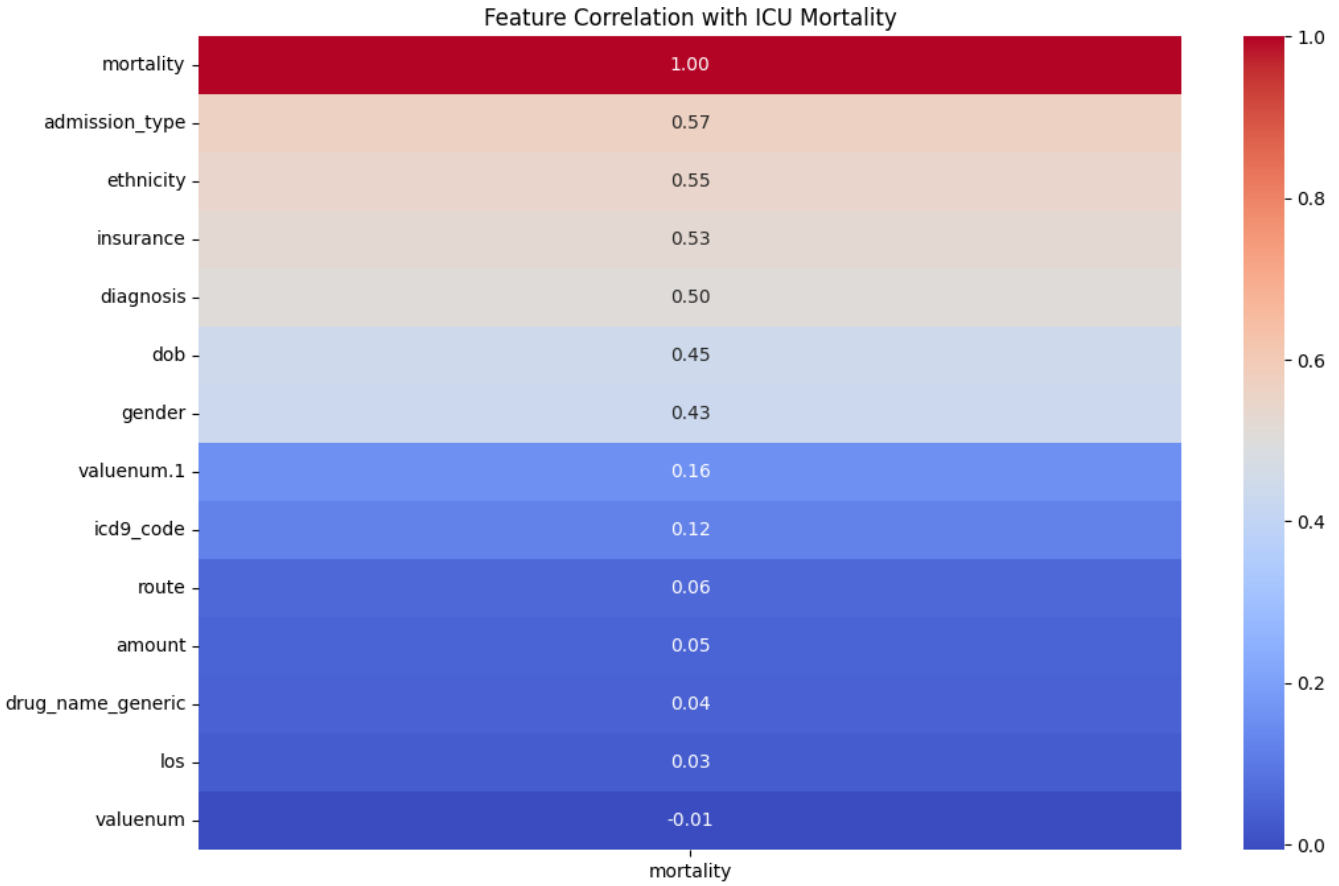

Figure 18. *The correlation matrix.*

6. Conclusion

6.1. Summary of Study

Interpretability has been one of the major issues facing the adoption of AI in high-stakes environments such as healthcare and finance. Many models that would have been useful in medical diagnosis and disease predictions are not explainable. This is because the models may have up to hundreds and thousands of neurons, each contributing to the decision-making process of the model. Much literature asserted that interpretability can be achieved in models with less complex structures; however, less complex structures are not so useful in everyday life scenarios. This research investigated that high accuracy can be achieved in healthcare prediction and diagnosis and at the same time achieve a detailed explanation of how the model makes predictions.

EHR data was used for this study. Findings reveal the strength of the LSTM model in processing time-series datasets. The study achieved 99.9% accuracy, 100% precision, and 99.8% recall. Besides, the SHAP model provided a detailed explanation of the LSTM model predictions.

This study demonstrates that SHAP can successfully enhance the interpretability of LSTM models applied to ICU mortality prediction. SHAP visualizations-including summary, force, and dependence plots-help explain both the overall importance of features and their individual effects on predictions. The results show a strong alignment with clinical knowledge, as features like diagnosis, age, and ICD codes had the highest impact on model output. Despite the complexity of LSTM models, interpretability was preserved through SHAP without a significant drop in predictive performance, achieving a balance between accuracy and transparency.

6.2. Contributions to Knowledge

The research presents a novel hybrid framework that integrate the LSTM deep learning model with SHAP interpretability to justify the possibility of achieving almost affect performance and obtain a detailed explanation of the func-

tionality of the model. The SHAP enhances transparency by breaking down the black-box nature of LSTM models and translating them into medically meaningful insights. It allows clinicians to understand not just what the model predicts, but why, facilitating informed trust and adoption in healthcare environments.

6.3. Limitations and Future Study

One of the major limitations of the study is that other interpretable techniques such as LIME, Random Forest, Integrated Gradients, etc can be used to explain the LSTM predictions. This may perhaps give a further explanation of what the SHAP model does not cover. Besides, other interpretable models can be used to evaluate the interpretability of the SHAP model, enhancing trust in the SHAP assertions. Besides, the study only used 22,625 instances. The EHR dataset has over 100,000 instances. More datasets can be used to train the LSTM model in the future to validate the outcomes obtained in this research. Additionally, real-time SHAP explanations could be embedded into clinical decision support tools to enable point-of-care interpretability, empowering healthcare professionals to make timely and well-informed interventions.

Abbreviation

LSTM	Long Short-Term Memory
SHAP	SHapley Additive exPlanations
EHR	Electronic Health Record
ICU Mortality	Intensive Care Unit Mortality
LIME	Local Interpretable Model-agnostic Explanations
CNN	Convolutional Neural Network
XAI	Explainable Artificial Intelligence

Author Contributions

Rotimi Philip Adebayo is the sole author. The author read and approved the final manuscript.

Funding

This work is not supported by any external funding.

Data Availability Statement

The MIMIC-III EHR dataset can be obtained from PhysioNet at https://physionet.org/content/mimiciii/1.4/

Conflicts of Interest

The author declare no conflict of interest.

Appendix

How SHAP Aligns with Clinical Knowledge

The SHAP chart in Figure 14 highlights clinically intuitive features such as diagnosis, date of birth (dob), and icd9_code as having the most significant impacts on the prediction of the model. This outcome supports established clinical understanding, where diagnostic codes and age are usually well-known features that determine the survival of patients in the ICU. Also, in the SHAP force plot for Instance 0 (Figure 16), features such as diagnosis = 52.0, dob = 48.5, and icd9_code = 214.0 predicts a lower mortality rate for that particular patient. The diagnosis, dob, and icd9_code present significant features that determine if a patient will survive or not. Such alignment between SHAP explanations and domain knowledge reinforces trust in the model's decision-making process, validating that the LSTM model is not learning spurious or medically irrelevant pattern.

How the Hybrid Model Improves Interpretability

By combining a deep learning model (LSTM) with SHAP explanations, the hybrid approach bridges the gap between performance and transparency. The LSTM model captures temporal patterns and interactions inherent in ICU data, while SHAP decodes these patterns into understandable visual insights. For instance, The LSTM model achieves nearly 100% accuracy even for unseen data. It perfectly predicted almost all True Positives and True Negatives correctly and only had 13 cases of False Negatives which it predicted wrongly. In the same vein, the SHAP perfectly provides a detailed interpretation of how the LSTM model makes predictions. The SHAP summary plot (Figure 15) ranks features by their average contribution to predictions, while the force plot (Figure 16) explains how each feature influenced a single prediction. This layered interpretability is especially useful in clinical contexts, where both the overall model behavior and individual predictions must be transparent for ethical and operational reasons.

Several studies have established an inverse relationship between accuracy and interpretability. While the trade-off evaluation is valid, this research presents a hybrid model that integrates the LSTM model with the SHAP model. This study shows that the trade-off can be reduced by using a hybrid approach that combines the LSTM model with SHAP. While LSTM provides a very strong predictive performance, SHAP provides a clear explanation of how the model makes decisions. The results of the study show that even with a complex model like LSTM, it is possible to provide a clear explanation of the model's behavior and decision-making through the hybrid model. This means the trade-off between accuracy and interpretability can be effectively managed, or even undermined when a post-hoc interpretability method like SHAP is applied.

Clinical Relevance and Insights

The SHAP model provides detailed insights into the pre-

diction process of the LSTM model. This is very important, especially in healthcare in that, medical practitioners can understand how the model works and juxtapose if the prediction aligns with established theories. It helps practitioners to know which variables are more critical and which of those are less critical.

References

[1] Herman, B. The Promise and Peril of Human Evaluation for Model Interpretability. *Arxiv*, Online; 2017.

[2] Lakkaraju, H., Bach, S. H., Leskovec, J. Interpretable Decision Sets: A Joint Framework for Description and Prediction. *In Proceedings of the 22nd ACM SIGKDD International Conference On Knowledge Discovery and Data Mining*. ACM, San Francisco; 2016, pp. 1675-1684. https://doi.org/10.1145/2939672.2939874

[3] Ribeiro, M. T., Singh, S., Guestrin, C. Anchors: High-Precision Model-Agnostic Explanations. *In Proceedings of the 32nd AAAI Conference on Artificial Intelligence (AAAI-18)*. Mcilraith, S. A., Weinberger, K. Q., Eds., AAAI Press, New Orleans, Louisiana, USA; 2018, Pp. 1527-1535.

[4] Angelino, E., Larus-Stone, N., Alabi, D., Seltzer, M., Rudin, C. Learning Certifiably Optimal Rule Lists. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, New York; 2017, pp. 35-44. https://doi.org/10.1145/3097983.3098047

[5] Yoon, C. H., Torrance, R., Scheinerman, N. Machine Learning In Medicine: Should The Pursuit of Enhanced Interpretability be Abandoned? *Journal of Medical Ethics*. 2022, 48: 581-585. https://doi.org/10.1136/Medethics-2020-107102

[6] Agarwal, C., Nguyen, A. Explaining Image Classifiers by Removing Input Features Using Generative Models. In *ACCV*. Springer, Cham; 2021, pp. 101-118. https://doi.org/10.1007/978-3-030-69544-6_7

[7] Doshi-Velez, F., Kim, B. Towards A Rigorous Science Of Interpretable Machine Learning. *Arxiv*, Ithaca, New York, USA; 2017.

[8] Miller, T. Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*. 2019, 267: 1-38. https://doi.org/10.1016/J.Artint.2018.07.007

[9] International Organization for Standardization (ISO). ISO/IEC TR 24028: 2020 — Information Technology — Artificial Intelligence — Overview of Trustworthiness in AI. Available from: https://www.iso.org/standard/77608.html (Accessed 10 July 2025).

[10] Markus, A. F., Kors, J. A., Rijnbeek, P. R. The Role of Explainability In Creating Trustworthy Artificial Intelligence For Health Care: A Comprehensive Survey Of The Terminology, Design Choices, And Evaluation Strategies. *Journal of Biomedical Informatics*. 2021, 113: 103655. https://doi.org/10.1016/j.jbi.2020.103655

[11] Esna-Ashari, M. Beyond The Black Box: Review Of Quantitative Metrics For Neural Network Interpretability And Their Practical Implications. *International Journal of Sustainable Applied Science and Engineering*. 2025, 2: 1-24.

[12] Yang, R. A., Jingyu, H., Zihao, L. C., Et Al. Interpretable Machine Learning for Weather and Climate Prediction: A Review. *Atmospheric Environment*. 2024, 120797. https://doi.org/10.1016/j.atmosenv.2024.120797

[13] Guangming, H., Yingya, L., Shoaib, J., Et Al. From Explainable To Interpretable Deep Learning For Natural Language Processing In Healthcare: How Far From Reality? *Computational and Structural Biotechnology Journal*. 2024. https://doi.org/10.1016/j.csbj.2024.05.004

[14] Giacobbe, D. R., Marelli, C., Guastavino, S., Et Al. Explainable and Interpretable Machine Learning for Antimicrobial Stewardship: Opportunities and Challenges. *Clinical Therapeutics*. 2024. https://doi.org/10.1016/j.clinthera.2024.02.010

[15] Kumar, A., Dikshit, S., Albuquerque, V. Explainable Artificial Intelligence for Sarcasm Detection in Dialogues. *Wireless Communications and Mobile Computing*. 2021, 1: 1-13. https://doi.org/10.1155/2021/2939334

[16] Ahmad, T., Katari, P., Kumar, A., Ravi, C., Shaik, M. Explainable AI: Interpreting Deep Learning Models for Decision Support. In *Advances in Deep Learning Techniques*. 2024, 4.

[17] Johansson, U., Sönströd, C., Norinder, U., Boström, H. Trade-Off between Accuracy and Interpretability for Predictive *In Silico* Modeling. *Future Medicinal Chemistry*. 2011, 3: 647-663. https://doi.org/10.4155/fmc.11.23

[18] Ribeiro, M. T., Singh, S., Guestrin, C. Why Should I Trust you? Explaining the Predictions of Any Classifier. *In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, San Francisco; 2016, pp. 1135-1144. https://doi.org/10.1145/2939672.293977

[19] Aldughay, B., Ashfaq, F., Jhanjhi, N. Z., Humayun, M. Explainable AI For Retinoblastoma Diagnosis: Interpreting Deep Learning Models with LIME And SHAP. *Artificial Intelligence in Medical Images*. 2023, 13: 1-13. https://doi.org/10.3390/diagnostics13111932

[20] Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., Elhadad, N. Intelligible Models for Healthcare: Predicting Pneumonia Risk and Hospital 30-Day Readmission. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 2015, Pp. 1721-1730. https://doi.org/10.1145/2783258.2788613

[21] Adebayo, R. P. Optimizing Food101 Classification with Transfer Learning: A Fine-Tuning Approach Using Efficientnetb0. *International Journal of Intelligent Information Systems*. 2024, 13: 59-77. https://doi.org/10.11648/j.ijiis.20241304.11

[22] Kanaung Academy. Understanding LSTM: A Deep Dive Into Its Inner Workings With Code Walkthrough. Available From: https://Medium.Com/@Kanaung.Academy/Understanding-Lstm-A-Deep-Dive-Into-Its-Inner-Workings-With-Code-Walkthrough-5337 (Accessed 3 August 2025).

[23] D2L Team. Dive Into Deep Learning: Long Short Term Memory. Available From: https://Classic.D2l.Ai/Chapter_Recurrent-Modern/Lstm.Html (Accessed 3 August 2025).

[24] Prasan, N. H. Activation Functions in Neural Networks. *Medium*. Available From: https://Medium.Com/@Prasannh/Activation-Functions-In-Neural-Networks-B79a2608a106 (Accessed 3 August 2025).

[25] Antonini, A. S., Tanzola, J., Asiain, L., Ferracutti, G. R., Castro, S. M., Bjerg, E. A., Ganuza, M. L. Machine Learning Model Interpretability Using SHAP Values: Application To Igneous Rock Classification Task. *Applied Computing and Geosciences*. 2024, 100178. https://doi.org/10.1016/j.acags.2024.100178

[26] Parisineni, S. R. A., Pal, M. Enhancing Trust and Interpretability Of Complex Machine Learning Models Using Local Interpretable Model-Agnostic SHAP Explanations. *International Journal of Data Science and Analytics*. 2024, 18: 457-466. https://doi.org/10.1007/s41060-023-00458-w

[27] Lipton, Z. C. The Mythos of Model Interpretability: In Machine Learning, The Concept of Interpretability is both Important and Slippery. *Queue*. 2018, 16: 31-57. https://doi.org/10.48550/arxiv.1606.03490

Biography

Rotimi Philip Adebayo is an Artificial Intelligence researcher and a student at ACETEL, National Open University of Nigeria (NOUN). He holds a Master's degree in Operations Research (2017) and a Master's degree in Computer Science (2012), both from the University of Lagos, Nigeria, and is currently pursuing a Master's degree in Artificial Intelligence. Recognized for his contributions to the field, he will commence a PhD program in Artificial Intelligence at the University of Portsmouth, UK, in 2025. His research interests include XAI, applied artificial intelligence, transfer learning, etc. He has authored several papers published in reputable international journals. In addition to his academic achievements, he is a professional data analyst and machine learning expert, holding Microsoft Azure Cloud certifications in Artificial Intelligence and Networking (CCNA).

Research Field

Rotimi Philip Adebayo: XAI, applied artificial intelligence, transfer learning, convolutional neural network, low-resource learning