IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 6 January 2024, accepted 6 February 2024, date of publication 13 February 2024, date of current version 20 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3365586

RESEARCH ARTICLE

Explainable Artificial Intelligence Model for Predictive Maintenance in Smart Agricultural Facilities

MELVIN KISTEN[1], ABSALOM EL-SHAMIR EZUGWU[1], AND MICHEAL O. OLUSANYA[2], (Member, IEEE)

[1]Unit for Data Science and Computing, North-West University, Potchefstroom 2520, South Africa
[2]Department of Computer Science and Information Technology, Sol Plaatje University, Kimberley 8300, South Africa

Corresponding author: Absalom El-Shamir Ezugwu (absalom.ezugwu@nwu.ac.za)

ABSTRACT Artificial Intelligence (AI) in Smart Agricultural Facilities (SAF) often lacks explainability, hindering farmers from taking full advantage of their capabilities. This study tackles this gap by introducing a model that combines eXplainable Artificial Intelligence (XAI), with Predictive Maintenance (PdM). The model aims to provide both predictive insights and explanations across four key dimensions, namely data, model, outcome, and end-user. This approach marks a shift in agricultural AI, reshaping how these technologies are understood and applied. The model outperforms related studies, showing quantifiable improvements. Specifically, the Long-Short-Term Memory (LSTM) classifier shows a 5.81% rise in accuracy. The eXtreme Gradient Boosting (XGBoost) classifier exhibits a 7.09% higher F1 score, 10.66% increased accuracy, and a 4.29% increase in Receiver Operating Characteristic-Area Under the Curve (ROC-AUC). These results could lead to more precise maintenance predictions in real-world settings. This study also provides insights into data purity, global and local explanations, and counterfactual scenarios for PdM in SAF. It advances AI by emphasising the importance of explainability beyond traditional accuracy metrics. The results confirm the superiority of the proposed model, marking a significant contribution to PdM in SAF. Moreover, this study promotes the understanding of AI in agriculture, emphasising explainability dimensions. Future research directions are advocated, including multi-modal data integration and implementing Human-in-the-Loop (HITL) systems aimed at improving the effectiveness of AI and addressing ethical concerns such as Fairness, Accountability, and Transparency (FAT) in agricultural AI applications.

INDEX TERMS Agriculture, smart agricultural facilities, predictive maintenance, machine learning, deep learning, explainable artificial intelligence.

LIST OF ABBREVIATIONS

Abbreviation	Definition
AI	Artificial Intelligence.
ANN	Artificial Neural Network.
AUC	Area Under the Curve.
CFE	Counterfactual Explanations.
CNN	Convolutional Neural Network.
CPU	Central Processing Unit.
CSIR	Council for Scientific and Industrial Research.
CSML	Cost-Sensitive Machine Learning.
DiCE	Diverse Counterfactual Explanations.
DL	Deep Learning.
DNN	Deep Neural Network.
DT	Decision Tree.
EL	Ensemble Learning.
ELI5	Explain Like I Am Five.
ELM	Extreme Learning Machine.
FAT	Fairness Accountability Transparency.
FN	False Negative.
FP	False Positive.
GB	Gradient Boosting.
GBDT	Gradient Boosting Decision Tree.
GBM	Gradient Boosting Model.
GDPR	General Data Protection Regulation.
GHz	Gigahertz.

The associate editor coordinating the review of this manuscript and approving it for publication was Tyson Brooks.

GPU	Graphics Processing Unit.
HITL	Human-in-the-Loop.
IAI	Interpretable Artificial Intelligence.
ID	Identifier.
LGBM	Light Gradient Boosting Model.
LIME	Local Interpretable Model-agnostic Explanations.
LOCF	Last Observation Carried Forward.
LR	Linear Regression.
LRP	Layer-wise Relevance Propagation.
LSTM	Long-Short-Term Memory.
LVQ	Learning Vector Quantisation.
MB	Megabyte.
ML	Machine Learning.
MLP	Multi-Layer Perceptron.
NN	Neural Network.
OC	One-Class.
OC-SVM	One-Class Support Vector Machine.
OS	Operating System.
PdM	Predictive Maintenance.
PPS	Predictive Power Score.
PRISMA	Preferred Reporting Items for Systematic Reviews and Meta-Analyses.
RAM	Random-Access Memory.
ReLU	Rectified Linear Unit.
RF	Random Forest.
RFECV	Recursive Feature Elimination with Cross-Validation.
RNN	Recurrent Neural Network.
ROC	Receiver Operating Characteristic.
ROC-AUC	Receiver Operating Characteristic-Area Under the Curve.
RTX	Ray Tracing Texel eXtreme.
RUL	Remaining Useful Life.
SAF	Smart Agricultural Facilities.
SHAP	SHapley Additive exPlanations.
SLR	Systematic Literature Review.
SRBD	Systematic Research on Big Data.
STD	Standard Deviation.
SVM	Support Vector Machine.
TANH	Hyperbolic Tangent.
TL	Transfer Learning.
TM	Trademark.
TN	True Negative.
TP	True Positive.
USA	United States of America.
VT	Variance Thresholding.
WoS	Web of Science.
XAI	eXplainable Artificial Intelligence.
XGB	eXtreme Gradient Boosting.

I. INTRODUCTION

Agricultural enterprises grapple with maintenance complexities and costly machinery [1], [2], [3], [4], [5], [6], [7]. These challenges threaten equipment upkeep and diminish productivity. Responding, farmers adopt Smart Agricultural Facilities (SAF), transforming farming practices with technologies like sensors and drones [8], [9]. These innovations not only streamline tasks such as soil classification [10], pest management [11], and water leakage detection [12], [13], but also enhance sustainability and efficiency. However, SAF systems, while transformative, are prone to malfunctions. The Predictive Maintenance (PdM) literature offers strategies to mitigate these problems, promoting operational efficiency, and reducing costs [14], thus sustaining and optimising SAF.

PdM in agriculture transcends mere data collection, revolutionising equipment upkeep. Farmers predict machine failures by leveraging data, enhancing efficiency, slashing costs, and boosting output [15]. However, the success of PdM and SAF is based on seamless data integration, user acceptance of new technologies, and strict privacy adherence [16], [17]; these factors determine the viability of these innovations in agriculture. With challenges in PdM and SAF, eXplainable Artificial Intelligence (XAI) becomes critical, especially under legal mandates like the General Data Protection Regulation (GDPR), demanding explainability in automated systems [18], [19]; this highlights the need for practical, user-friendly, and legally compliant AI models [20]; this study probes a fundamental question: How does one trust AI predictions in SAF without explanations? It explores XAI model development, focusing on PdM strategies [21] and AI explainability methods [22].

Recent AI models for PdM in SAF often obscure their decision-making process, a significant shortfall [23]; this opaqueness, exacerbated by literature focusing narrowly on accuracy and F1 scores without detailed train and test results, diminishes transparency and robustness. Additionally, AI's role in agriculture, as either a farmer's aid or replacement, remains unclear [24]. Consequently, farmers find current AI models in agriculture challenging to grasp, limiting their utility; this gap calls for XAI to improve the understanding and usability of AI in PdM and SAF [25], [26], [27], [28]. Therefore, this study proposes a model blending XAI with PdM in SAF to improve transparency and practicality.

Although some studies on AI explainability adopt a broad, universal approach [29], this study seeks a more detailed, flexible explainability method, tailored to varied stakeholder needs. It focuses on matching explanation depth to the unique needs of diverse users, like AI experts and farmers; this approach recognises that explanation purposes differ, requiring a more customised framework. Specifically, this study highlights the importance of explanations that provide scientific insight and meet stakeholders' internal needs, aiming for a more refined use of AI in agricultural PdM.

Embarking on this study involves conducting a thorough Systematic Literature Review (SLR) with bibliometric analysis to understand the current state of AI models, honing in on XAI. It involved delving into the literature on Deep Learning (DL) and Machine Learning (ML) in PdM, emphasising

XAI's importance due to the opaque decision-making of AI models in SAF [23]. These explorations shape the study's objectives and influence the development of a combined model, highlighting XAI's role in agricultural PdM. The study evaluates four DL and seven ML algorithms, selecting the top-performing ones based on the test F1 score for explainability and for comparative analysis. The research aims to (1) predict maintenance needs and (2) provide explanations using XAI in SAF. To this end, the study focuses on: (i) Assessing AI models' status and the need for explainability through SLR and bibliometric analysis. (ii) Develop DL and ML models for PdM. (iii) Evaluation of the effectiveness of these models. (iv) Elucidating the rationale behind the AI model's predictions. The key contributions of this study are summarised as follows:

1. It pioneers integrating XAI with PdM in SAF, focusing on four explainability dimensions: (1) data, (2) model, (3) outcome, and (4) end-user.
2. Conducts an SLR on the current state of AI models, highlighting the necessity of XAI.
3. Presents transparent training and testing outcomes using the more stringent Receiver Operating Characteristic-Area Under the Curve (ROC-AUC) metric, underscoring the importance of XAI beyond just accuracy.
4. Adapts design principles from XAI and PdM for the specific context of SAF.
5. Strengthens the theoretical underpinnings and future trajectories of XAI in PdM for SAF.

II. LITERATURE REVIEW

This study undertook an extensive literature review to analyse advanced AI-driven PdM techniques critically. It aimed to identify and assess XAI models for PdM, focusing on their role in enhancing maintenance practices. The study also aimed to shed light on the strengths, limitations, and practicality of various XAI models in PdM. Evaluating and synthesising these findings aimed to provide a comprehensive overview of the current state and future direction of XAI models in agricultural PdM.

A. PREDICTIVE MAINTENANCE

This review uncovered three PdM approaches: (1) anomaly detection, (2) prognostics, and (3) diagnostics [21]. Anomaly detection is akin to spotting the unusual in data. Prognostics predict future system performance, while diagnostics identify current issues through performance analysis. Among the reviewed literature, eleven studies focused on prognostics [30], [31], [32], [33], [34], [35], [36], [37], [38], [39], [40], three on anomaly detection [41], [42], [43], and two on both prognostics and diagnostics [40], [44]. Notably, none focused solely on diagnostics, revealing a significant research gap. Future studies should explore how anomaly detection and prognostics can lead to effective diagnostics, thereby improving the robustness and efficiency of PdM in SAF.

B. DEEP AND MACHINE LEARNING FOR PREDICTIVE MAINTENANCE

For prognostics, the Recurrent Neural Networks (RNNs) and Long-Short-Term Memory (LSTM) stand out for their precision; they achieved a notable 90.07% accuracy [33]. Similarly, in predicting Remaining Useful Life (RUL), boundaries were pushed with Bidirectional Recurrent Neural Networks (Bi-RNN) and LSTMs, attaining a 96.15% precision [30]. LSTMs play a pioneering role in anomaly detection, complemented by One-Class Support Vector Machines (OC-SVM), which significantly reduce false alarms [38]. However, OC-SVMs struggle with supervised problems.

An alternative study that used Random Forest (RF) in prognostics also incorporated AutoML. The Random Forest (RF), showed versatility, especially in component-level analysis [32], [36]. However, AutoML's generalist approach, while democratising ML, hinders model optimisation [45]. Ensemble Learning (EL) offered diverse algorithmic solutions and found utility in the prognostics realm of manufacturing industries [36]. Despite their complexity, other methods like Balanced K-Star, Multi-Layer Perceptron (MLP), Extreme Learning Machine (ELM), and Transfer Learning (TL) provide alternative approaches, as do Deep Convolutional AutoEncoders [34], [35], [40], [46]. Diagnostics in PdM have been less explored, with only a few studies touching on it [40]. In general, PdM combines a variety of methodologies, each with unique strengths and challenges, necessitating ongoing critical evaluation.

C. EXPLAINABLE ARTIFICIAL INTELLIGENCE

In the fast-paced era of DL and ML, sophisticated model deployment now pervades sectors like healthcare, finance, and agriculture. However, the intricate nature of these models often clouds their decision-making processes, raising concerns about transparency [23]. This opacity has catalysed the emergence of explainability in DL and ML, a concept that transcends mere transparency. Explainability entails dissecting the complexities of the DL and ML models to render their decision-making understandable, catering to both experts and non-experts. Crucially, explainability embodies a spectrum of facets, each pivotal to demystifying and ensuring the reliability of the DL models.

1) DIMENSIONS OF EXPLAINABILITY

Four explainability dimensions were extracted from the reviewed literature: (1) data, (2) model, (3) outcome, and (4) end-user. The ''data dimension'' delves into the data's limitations and potential [22]. However, most studies overlooked these aspects, failing to assess whether the data could support the insights sought; this oversight calls for more in-depth research on the data capabilities for PdM in SAF. The ''model dimension'' explores how input data influences model predictions [22]. Often, researchers assume feature independence, a notion prone to bias. Despite this, most studies incorporated this dimension, with a few focusing

on both model and ''outcome dimensions'' [30], [31], [32], [36], [39]. The review revealed only two studies dedicated to outcome explainability [34], [37], indicating a research gap in understanding the reasoning behind single-instance AI model predictions. Addressing this can enhance transparency and decision-making in AI models. The ''end-user'' dimension, which tailors explanations to non-technical users [47], remained unexplored in the reviewed literature, signalling a need for research that makes AI understandable to a broader audience.

2) APPROACHES TO EXPLAINABILITY

Six explainability approaches emerged from the literature [48]: local explainability, global explainability, model-specific, model-agnostic, model-centric, and data-centric approaches. Local explainability clarifies individual predictions, while global explainability unveils overall model behaviour. Two studies tackled both local and global explainability [33], [41], but none solely focused on global explainability, indicating a research gap. Thirteen studies explored local explainability alone [30], [31], [32], [34], [35], [36], [37], [38], [39], [40], [42], [43], [44].

Model-specific approaches are confined to specific AI models, whereas model-agnostic strategies apply universally. Ten studies used model-agnostic methods [30], [31], [32], [34], [36], [38], [40], [42], [43], [44], while three used model-specific approaches [33], [35], [37], and only two studies harnessed both [39], [41]. Model-centric approaches analyse input-output relationships within models, while data-centric approaches focus on data quality and relevance [47]. All reviewed studies focused on model-centric approaches, leaving data-centric strategies largely unexplored, thus highlighting a significant research opportunity.

D. EXPLAINABLE ARTIFICIAL INTELLIGENCE FOR PREDICTIVE MAINTENANCE

SHapley Additive exPlanations (SHAP) stood out, despite complexities [30], [32], [36], [38], [39], [40], [42], [44]. SHAP was used to clarify feature impacts for false alarm predictions [38]; its diagnostic application helped in model interpretation [40], [44]. Local Interpretable Model-agnostic Explanations (LIME), versatile in various AI models, excelled in specific, localised predictions, as demonstrated in transportation anomaly detection [41]. However, LIME's local focus limits a broader understanding of the model.

Layer-wise Relevance Propagation (LRP), suited for DL, offered detailed insights into prediction influences [33]. LRP's effectiveness was notable, but model-specific. A comparison of LIME, SHAP, and Explain Like I Am Five (ELI5) revealed varied feature attributions and efficiency [39]. LIME was shown to be efficient; ELI5 provided more intuitive explanations, but lacked model-agnostic versatility. Counterfactual Explanations (CFE) have gained popularity for enhancing AI acceptability, especially among non-experts [34], [36]. Consequently, while XAI for PdM presents various tools, challenges in complexity, accessibility, and applicability, these highlight the ongoing quest for a balance between technical depth and user-friendly explanations.

TABLE 1. Databases and search strings.

Database	Search string
Web of Science (WoS)	(TS=(explain* OR interpret* OR xai OR iai) AND TS=(ai OR ml) AND TS=(maintenance OR detect* OR prognos* OR diagnos*)) AND (PY==("2012" OR "2013" OR "2014" OR "2015" OR "2016" OR "2017" OR "2018" OR "2019" OR "2020" OR "2021" OR "2022") AND DT==("ARTICLE") AND LA==("ENGLISH") AND SJ==("COMPUTER SCIENCE"))
Scopus	(TITLE-ABS-KEY (explain* OR interpret* OR xai OR iai) AND TITLE-ABS-KEY (ai OR ml) AND TITLE-ABS-KEY (maintenance OR detect* OR prognos* OR diagnos*)) AND PUBYEAR > 2011 AND PUBYEAR < 2023 AND (LIMIT-TO (SUBJAREA , "COMP")) AND (LIMIT-TO (DOCTYPE , "ar")) AND (LIMIT-TO (LANGUAGE , "English")) AND (LIMIT-TO (SRCTYPE , "j"))

TABLE 2. Search parameters.

Parameter	Value
Timespan	2012-2022
Language	English
Subject area	Computer Science
Document types	Article
Sources	Journals

E. BIBLIOMETRIC ANALYSIS

This study undertook a detailed bibliometric analysis emphasising transparency in AI. The methodology encompassed data processes, descriptive statistics, keyword analysis, knowledge synthesis, and exploration of conceptual and social structures. Searches focused on ''Titles'', ''Abstracts'', and ''Subject Headings'', covering literature from 2012 to 2022, excluding the incomplete year 2023. In the databases, ''Keywords Plus'' and ''Keywords'' served as Subject Headings, enabling a thorough search through synonyms, domain-specific terms, and relevant keywords [49]; this approach helped identify literature pertinent to the research topic. Table 1 details the specific search strings used in the Web of Science (WoS) and Scopus databases. Also, note the parameters selected in the search strings in Table 2.

Fig. 1 depicts a Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) flow diagram [50], outlining the data collection, screening, and analysis steps and how resources were chosen or excluded for the review. The ''Identification'' stage involved gathering 1711 records from two academic databases (WoS and Scopus), with

361 duplicates removed. Screening assessed each record's relevance, leading to the exclusion of 51 preprints and non-journal records, based on criteria in Table 2. The ''Included'' stage refers to the final selection of 759 records for the review.

1) REVIEW METHODOLOGY

The review involved a rigorous bibliometric analysis, sourcing data from the WoS and Scopus databases on 30 June 2023. A uniform search string on both platforms yielded 759 publications for this review. Data from WoS and Scopus were merged using the Bibiometrix R software package [51], employing the ''mergeDbSources'' function for integration and the ''remove.duplicated'' feature to ensure no duplication. Data pre-processing, crucial for accuracy, involved filtering terms like ''na'', ''n/a'', ''n.a'', and ''0'', and managing synonyms to maintain data consistency [52]. The study conducted a comprehensive bibliometric and critical analysis using RStudio, enhanced by the Bibliometrix package [51], streamlining the review [53].

2) DESCRIPTIVE STATISTICS

The bibliometric analysis included 759 journal articles, with an average age of 2.16 years and an average citation count of 14.24 per document. The yearly expansion rate stood at 44.85%, encompassing 35274 references. The analysis revealed 3878 instances of ''Keywords Plus'' [49] and 2564 instances of ''Author Keywords''. Table 3 provides an overview of the final dataset used in this bibliometric analysis.

3) KEYWORD ANALYSIS

Keywords served as vital links, guiding researchers to relevant materials through academic databases, and embodied the core ideas of a research topic; their absence makes finding pertinent documents challenging. Author keywords, self-selected by authors, represent the essence of their work. In this review, Fig. 2 shows the top 10 author keywords in the dataset. ''machine learning'' led with 212 instances, reflecting its prominence in XAI and PdM research. Following were ''explainable artificial intelligence'' and ''deep learning'', with 208 and 145 occurrences, respectively. These author-provided keywords efficiently navigate the knowledge landscape of scientific fields.

4) KNOWLEDGE SYNTHESIS AND CONCEPTUAL STRUCTURE

Science mapping, a pursuit to unravel the network within evolving scientific knowledge [54], aimed to uncover the structure and dynamics of scientific research. It offered a statistical lens to delve into scientific domains, highlighting key themes and developments. Using the Walktrap algorithm [55], the interactive co-occurrence grid visualised in Fig. 3, standardised associations, and linked words within a unified document; this approach illuminated topics and insights within a research field and mapped the evolutionary path of studies over time. Fig. 3's network portrays the progression of XAI in PdM research, demonstrating the frequent co-occurrence of ''explainable artificial intelligence'' and ''deep learning'', underscoring the growing emphasis on integrating explainability into DL algorithms.

TABLE 3. Overview of the final dataset (2012 – 2022).

Description	Results
Timespan	2012:2022
Sources (Journals, Books, among others.)	298
Documents	759
Annual growth rate%	44.85
Average age of the document	2.16
Average citations per doc	14.24
References	35274
Document contents	
Keywords Plus	3878
Author's Keywords	2564
Authors	
Authors	3559
Authors of single-authored docs	24
Authors' Collaboration	
Single-authored docs	24
Co-Authors per Doc	5.43
International co-authorships%	20.42
Document types	
Article	759

5) SOCIAL STRUCTURE

Fig. 4 highlighted Wang Y and Zhang Y as prominent contributors with significant collaboration scores in the field. Interestingly, their collaborative efforts were not directed toward each other.

Fig. 5 shows that prominent institutions like the University of Florida, Tsinghua University, and Central South University held significant ranks in collaboration. However, their collaborative efforts remain separate, without intersections.

Similarly, Fig. 6 revealed that the United States of America (USA) predominantly collaborated with China and India in its research efforts. Following this detailed bibliometric analysis, the next section presents the critical analysis and identified research gaps, elucidating this study's rationale.

F. CRITICAL ANALYSIS

Analysing research gaps is crucial to highlighting what is lacking in the literature and driving this research effort. It reveals sectors that need XAI and influences the methodologies chosen. Bibliometric analysis also leads to the need for XAI. The review identified three sectors using PdM with XAI: (a) transportation, (b) manufacturing, and (c) smart grids. Among these, manufacturing garnered the attention of

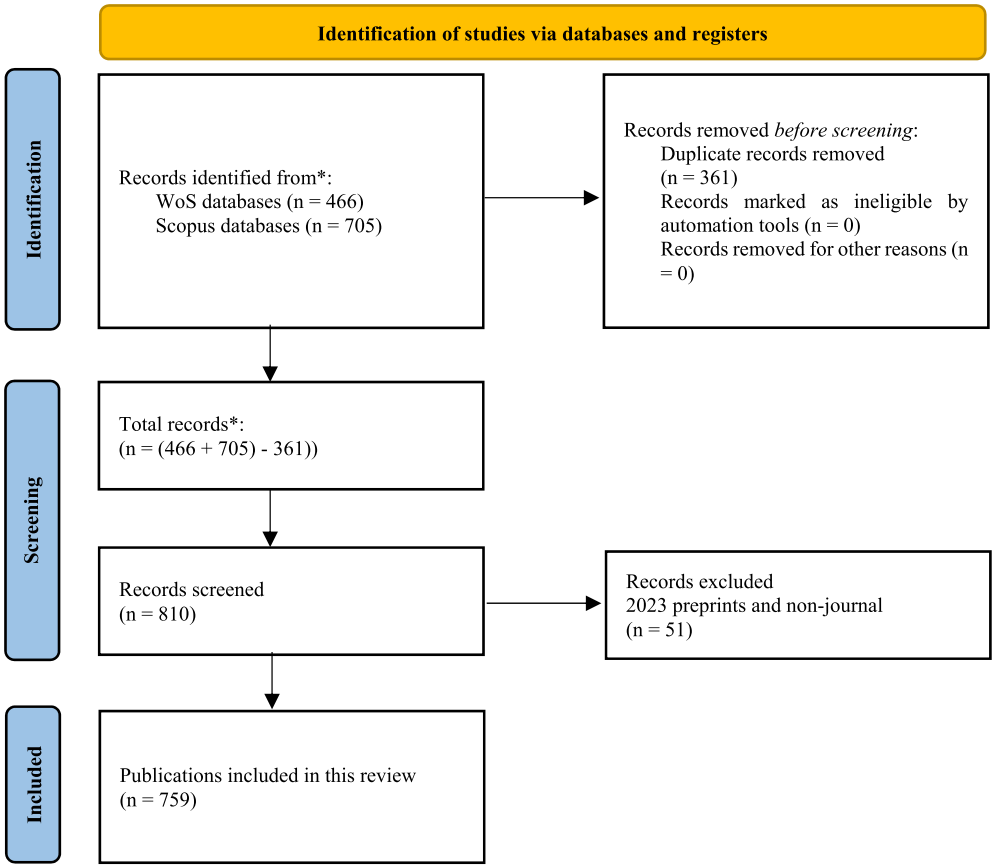


FIGURE 1. Summary of the data extraction and screening process.

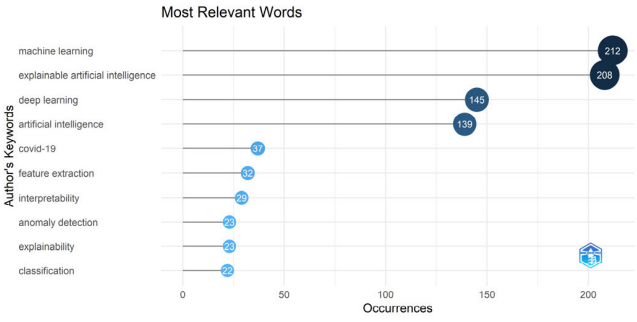


FIGURE 2. Top 10 most frequent author keywords in the dataset.

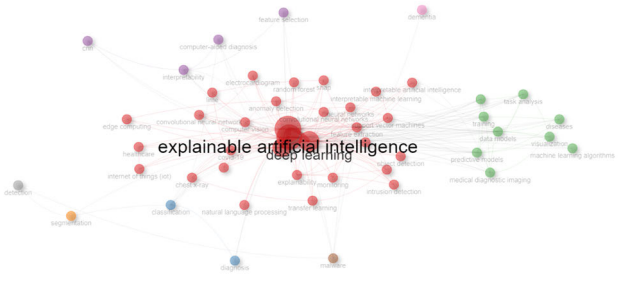


FIGURE 3. Co-occurrence network.

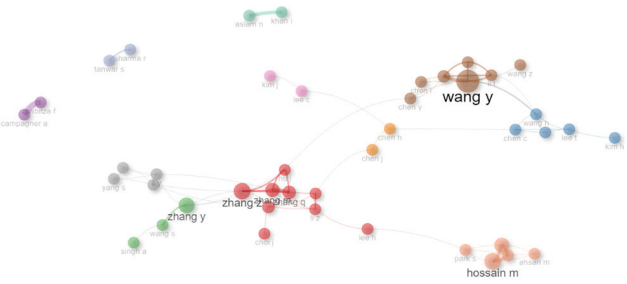


FIGURE 4. Collaboration network by author.

FIGURE 5. Collaboration network by institution.

twelve studies [31], [32], [33], [34], [35], [36], [37], [38], [40], [42], [43], [44], while transportation claimed two [30], [41], and smart grids only one [39].

Various state-of-the-art models, like Ensemble Learning (EL) [36], [44], and Deep Learning [30], [33], [37], were used in PdM; this study will use similar algorithms, aiming for comparable results and employing metrics like F1 score,

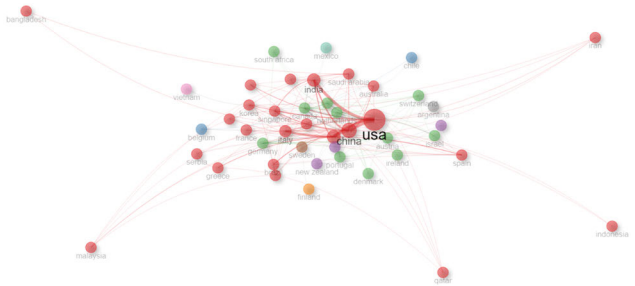


FIGURE 6. Collaboration network by country.

accuracy, and ROC-AUC for evaluation [30], [32], [33], [44], [56], [57]. Also, due to the prevalent usage of the test F1 score and for comparative analysis, it will be used to select the most outstanding DL and ML algorithms on each dataset, which will then be used for the explainability stage.

Popular explainability techniques included ELI5, LIME, SHAP, and LRP. Where ELI5 focused on feature impact, while [39] SHAP provided global explanations [30], [32], [36], [38], [39], [40], [42], [44] and LRP suits DL models [33], [41]. Although CFEs showed their application in the outcome dimension [34], [36], this study will embrace them to address the end-user dimension, offering choices for desired outcomes. Unfortunately, most studies do not address the data dimension's explainability [22], a gap that this research aimed to fill.

The bibliometric analysis revealed a prevailing trend of AI models being used in PdM, substantiating the literature review finding [31], [32], [36], [38], [39], [40], [42], [44] (refer to Fig. 2). Additionally, it underscored the prevalent use of XAI in DL, reinforcing the imperative need for explainability in DL (refer to Fig. 3). The social structure also showed an increasing demand for collaborative efforts to advance XAI in PdM research (refer to Fig. 4, 5, and 6). Notable gaps include data, outcome, end-user explainability, and the need for global explainability combined with model-specific and model-agnostic methods. There is also a gap in integrating anomaly detection and diagnostics, and a lack of focus on diagnostics alone. The intersection of XAI and PdM in SAF was under-researched, impeding farmers' understanding and usage of AI. Therefore, this study proposed a combined model that merges XAI and PdM to (1) predict maintenance needs and (2) provide explanations for the predictions made by the model, trained on a time series dataset of maintenance records, sensor readings dataset, for machine and water pump statuses. The proposed AI model predicted potential failures (prognostics), highlighting their root causes by component level (diagnostics). Also, the model addressed four dimensions: (1) data, (2) model, (3) outcome, and (4) end-user [22], [47].

III. EXPERIMENTAL DESIGN

This study aimed to predict maintenance needs and provide explanations using XAI. Also, it outlines the experimental design, covering the datasets, research design, system and parameter configurations. To achieve its objectives, the study conducted various experiments.

TABLE 4. Telemetry attributes.

No	Attribute	Description
1	datetime	When the reading was recorded
2	machineID	Unique ID for each machine
3	volt	The potential difference between two points in an electrical circuit
4	rotate	The speed at which the pump motor rotates
5	pressure	The effect at which water is carried inside the pipe
6	vibration	Rapid back-and-forth motion of the water pump
7	errorID	Five unique error types that a machine can have ['error1', 'error4', 'error3', 'error5', 'error2']
8	comp	There are four types of components for which a machine goes for maintenance ['comp2', 'comp1', 'comp4', 'comp3']
9	model	Four unique machine models ['model3', 'model4', 'model2', 'model1']
10	age	Age of machine
11	failure	Failure due to a specific component ['comp1', 'comp3', 'comp4', 'comp2']

A. DATA

Additionally, this study used open-source data to experiment with and assess the proposed AI model, enhancing reproducibility. It used the Telemetry for Predictive Maintenance dataset [58] with 11 attributes, and the Pump Sensor dataset [59] with 54 attributes. The Telemetry data covered hourly intervals from 06:00:00 on the first day of 2015 to the same time on the first day of 2016. The Pump Sensor data spanned from midnight, on 1 April 2018, to one minute before midnight on 31 August 2018, recording every minute and included data from 52 sensor units before pre-processing. The attributes and definitions within the Telemetry Predictive Maintenance dataset are shown in Table 4. The attribute and meaning of the Pump Sensor dataset are shown in Table 5.

After pre-processing, the Telemetry for Predictive Maintenance dataset had 292019 observations with 19 unique attributes, while the Pump Sensor dataset comprised 218880 observations with 4 attributes. Both datasets revealed a significant class imbalance. In the Telemetry dataset, normal (class none) instances exceeded failure instances (comp1, comp2, comp3, comp4) by 98%. Similarly, in the Pump Sensor dataset, normal status instances surpassed abnormal (recovering, broken) by 93%. Cost-Sensitive Machine

TABLE 5. Pump sensor attributes.

No	Attribute	Description
1	datetime	When the reading was recorded
2	sensor_n	Specific sensor and reading (n = 52)
3	pump_status	Three pump states ['normal', 'broken', 'recovering']

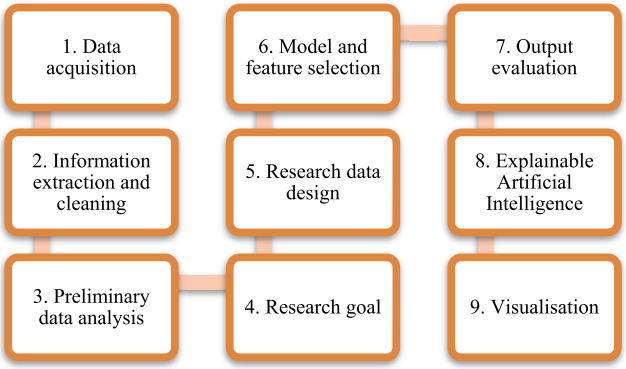


FIGURE 7. Modified SRBD stages.

Learning (CSML) was applied before the modelling stage to address these imbalances, ensuring fair representation and accuracy across different classes.

B. RESEARCH DESIGN

Influenced by the objectives of this study, the research design adopted the Systematic Research on Big Data (SRBD) methodology [60]. Despite critiques about SRBD's lack of role clarity [61], its data-driven, agile nature suited this academic research, enhancing reproducibility. SRBD typically involves seven stages: information extraction and cleaning, preliminary analysis, defining research goals, data design, model and feature selection, output evaluation, and visualisation; this study expanded to nine stages, each tailored to specific objectives and shown in Fig. 7. Stage 1, data acquisition, was pivotal for later stages. Stage 2 focused on XAI, deciphering predictions made by the model. A thorough SLR with bibliometric analysis commenced, addressing the (i) landscape of AI models and the need for explainability. Stages 2 to 6 concentrated on (ii) developing an AI model for PdM, stage 7 (iii) evaluating it, and the final stages, 8 and 9, aimed at (iv) providing explanations for the model's predictions, meeting all research objectives comprehensively.

In this study, data acquisition denoted the sourced machine and irrigation system data, analysing component-specific failures and pump failures. Information extraction and cleaning involves data pre-processing, rendering the data fit for subsequent stages [52]; this procedure encompassed (1) data cleaning, (2) data formatting, (3) feature engineering, and (4) data integration.

Additional information on these are provided as follows: (i) Data cleaning. The study employed "forward fill" for missing values, avoiding data leakage, also known as Last Observation Carried Forward (LOCF), which replaces missing values with their preceding present value, honouring the dataset's temporal order. Moreover, this method is effective when working with time series data [62], [63]. (ii) Data formatting deals with categorised variables for efficient processing. (iii) Feature engineering was used to extract features: (a) dateTime features, (b) lag features, (c) window features, and (d) periodic cyclic features. (iv) Data integration relates to merged multiple files in the Telemetry for Predictive Maintenance dataset [58]. Moreover, preliminary data analysis is focused on data analysis.

The research goal is aimed at predicting maintenance needs and providing explanations using XAI. Research data design involves structured data for AI modelling, ensuring temporal order during encoding and scaling, and split data into 70% training and 30% testing sets. In terms of the model and feature selection, the selected models and features were based on the literature review, using the test F1 score as the benchmark and for comparative analysis [32], [44]. The following were implemented in terms of the feature selection process: (1) Variance Thresholding (VT), (2) pairwise correlation, (3) Recursive Feature Elimination with Cross-Validation (RFECV), and (4) Boruta-SHAP.

For the output evaluation, the F1 score, accuracy, ROC-AUC, and class weights were used for performance evaluation, which were widely used in academic literature and those particularly tailored for dealing with imbalanced data [32], [33], [44], [56], [57]. A classifier could land within any category of the confusion matrix: *TP, TN, FP, or FN*. Where: TP denotes True Positive (correctly identified), TN denotes True Negative (correctly rejected), FP denotes False Positive (incorrectly identified), and FN denotes False Negative (incorrectly rejected). Moreover, the performance evaluation of a classification algorithm hinges on using a confusion matrix; this matrix presents correct classifications compared with erroneous ones, all categorised per class [64], [65]. Furthermore, the F1 score embodies the harmonic average, combining recall and precision; the former gauges the accurate prediction of true positives, whilst the latter quantifies the misrepresentation of the positive class. A mathematical definition of the F1 score metric emerges in Equation (3). Note that the precision and recall must first be computed:

$$Precision = \frac{TP}{TP + FP} \tag{1}$$

$$Recall = \frac{TP}{TP + FN} \tag{2}$$

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{3}$$

In this context, the accuracy shows the proportion of accurate classifications relative to the total classifications. A mathematical definition of the accuracy metric emerges

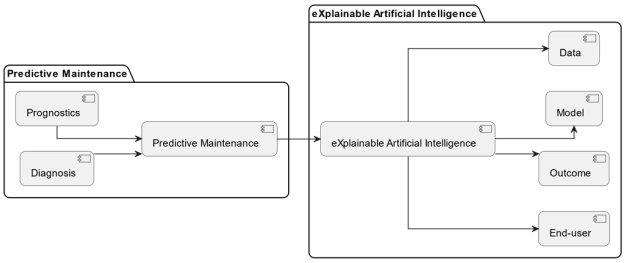


FIGURE 8. PdM and XAI illustration.

in Equation (4):

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

The ROC visually shows model effectiveness, while the AUC numerically grades it from 0 to 1 [66], [67]. The AUC's versatility in appraising performance across various thresholds renders it crucial, mainly when classes differ in significance or risk. This study used the Trapezoidal rule for AUC calculation, among other methods like Riemann Sums and Simpson's rule; this approach is particularly relevant here, as accurate failure prediction is deemed more critical than correctly identifying non-failures, reflecting the unequal importance of different classes in this context. More so, the design involving eXplainable Artificial Intelligence is targeted toward transparency in predictions across four dimensions: (1) data, (2) model, (3) outcome, and (4) end-user [22], [47].

Summarily, this research advanced a unified model that fused XAI with PdM to (1) predict maintenance needs and (2) furnish explanations for the model's predictions, as seen in Fig. 8. Moreover, trained on a time series dataset containing maintenance logs, sensor data, and machine and water pump statuses, the proposed AI model predicts potential failures (prognostics), elucidating their underlying causes by component level (diagnostics) and addressing explainability through four dimensions: (1) data, (2) model, (3) outcome, and (4) the end-user [22], [47]. Finally, the study explored the option of using SHAP plots and CFE tables for intuitive explanation and visualisation presentation.

C. SYSTEM AND PARAMETER CONFIGURATION

Research experiments were done using an Intel(R) Core (TM) i5.10500H Central Processing Unit (CPU) at 2.50 GHz speed, and 16174 MB RAM, running on Windows 11 Pro 64-bit Operating System (OS) with a 6009 MB Nvidia GeForce RTX 3060 laptop Graphics Processing Unit (GPU). The Python 3.8.17 language underpinned the programming. The research detailed configurations for top-performing DL and ML algorithms per dataset. It also outlined parameter settings for the four explainability dimensions: (1) data, (2) model, (3) outcome, and (4) end-user [22], [47]. The sequence began with DL configurations for the Telemetry for Predictive Maintenance dataset, followed by those for explainability.

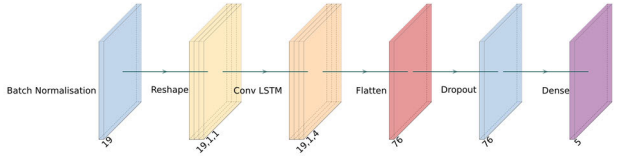


FIGURE 9. DL - Convolutional long-short-term memory neural network parameters (telemetry for predictive maintenance).

It then presented DL configurations for the Pump Sensor dataset and their explainability aspects, with a similar pattern for ML configurations.

1) TELEMETRY FOR PREDICTIVE MAINTENANCE DATASET CONFIGURATIONS

This study divided its dataset, allocating 70% to training and 30% to testing. A consistent random state, fixed at 777, was applied throughout all experiments. The model input comprised 19 features, with output targeting classification across 5 classes.

a: DEEP LEARNING (TELEMETRY FOR PREDICTIVE MAINTENANCE)

The Convolutional Long-Short-Term Memory (LSTM) Neural Network's parameters, displayed in Fig. 9, involve a sequence where input data are first transformed, normalised, and scaled; this ensures efficient training. The data are then reshaped into a (19, 1, 1) configuration, followed by a 1D Convolutional LSTM with four output channels. Subsequently, the input is flattened into a vector, retaining its initial form. To enhance regularisation, a certain percentage of inputs are randomly dropped. The final step involves a fully connected layer with 5 output units, facilitating the classification into 5 classes.

i) DATA EXPLAINABILITY (TELEMETRY FOR PREDICTIVE MAINTENANCE)

Table 6 shows the parameters for the Deepchecks data explanations.

TABLE 6. Deepchecks data explainability parameters (telemetry for predictive maintenance).

Parameter	Value	Description
n_samples	len(dataset)//2	The number of samples to use in the data integrity check
timeout	600	The maximum amount of time (in seconds) to allow for the check
n_top_columns	50	The number of top columns to include in the check
n_to_show	50	The number of results to show

ii) MODEL EXPLAINABILITY (TELEMETRY FOR PREDICTIVE MAINTENANCE)

Table 7 shows the parameters for the DL SHAP model explanations.

TABLE 7. DL - SHAP model explainability parameters (telemetry for predictive maintenance).

Parameter	Value	Description
sample_size	25	Number of samples taken from the training data for SHAP analysis
X_sub	df_train	Subset of the training data (excluding the target variable) used for SHAP analysis
explainer	SHAP explainer object	An object that can calculate SHAP values, created using the model's prediction function and sampled data
shap_values_multiple	df_test[0:sample_size]	The SHAP values for the test data
class_indices	All five classes	List of class indices for which SHAP summary plots will be generated
class_label	['comp1', 'comp2', 'comp3', 'comp4', 'none']	The label of the class corresponding to the current class index

iii) OUTCOME EXPLAINABILITY (TELEMETRY FOR PREDICTIVE MAINTENANCE)

Table 8 shows the parameters for the explanations of the DL SHAP outcomes.

TABLE 8. DL - SHAP Outcome explainability parameters (telemetry for predictive maintenance).

Parameter	Value	Description
sample_size	25	Number of samples taken from the training data for SHAP analysis
X_sub	df_train	Subset of the training data (excluding the target variable) used for SHAP analysis
explainer	SHAP explainer object	An object that can calculate SHAP values, created using the model's prediction function and sampled data
shap_values	df_test[0:1]	The SHAP values for the test data
class_indices	All five classes	List of class indices for which SHAP summary plots will be generated
class_label	['comp1', 'comp2', 'comp3', 'comp4', 'none']	The label of the class corresponding to the current class index

iv) END-USER EXPLAINABILITY (TELEMETRY FOR PREDICTIVE MAINTENANCE)

Table 9 shows the parameters for the DL DiCE end-user explanations.

TABLE 9. DL - DiCE end-user explainability parameters (telemetry for predictive maintenance).

Parameter	Value	Description
data_object	dice_ml.Data object	DiCE data object prepared using the input dataset
backend	'TF'+tf.__version__[0]	TensorFlow backend version used (e.g., 'TF2' for TensorFlow version 2.x)
model_object	dice_ml.Model object	DiCE model object created using the TensorFlow model
explainer	dice_ml.Dice object	DiCE explanation object instantiated with the data object, model object, and the method set to 'random'
desired_classes	All other classes	List of desired classes for which counterfactuals are generated
test_query	DataFrame object	Subset of a dataset (excluding the target variable) representing a test query for which counterfactuals are generated
total_CFs	5	Total number of counterfactual instances to generate for each desired class
features_to_vary	"all"	Specification of features to vary during counterfactual generation
proximity_weight	1.5	Weight is assigned to proximity in the counterfactual generation process. Feature-wise distance from the original input
diversity_weight	1.0	The weight assigned to diversity in the counterfactual generation process. Feature-wise distance between each counterfactual pair
stopping_threshold	0.5	Threshold for stopping the counterfactual generation process

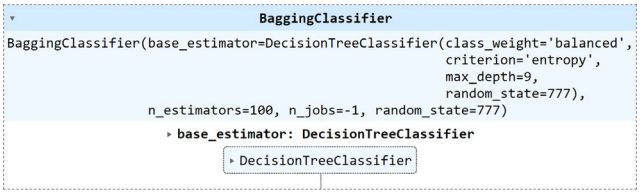


FIGURE 10. ML - Bagging classifier parameters (telemetry for predictive maintenance).

b: MACHINE LEARNING (TELEMETRY FOR PREDICTIVE MAINTENANCE)

Bagging uses an ensemble of Decision Trees (DT) (Fig. 10). Weights find themselves assigned to classes, harmonising model performance. The entropy criterion comes into play for DT node splitting. The maximum depth of the DT receives specification. The ensemble boasts 100 DT estimators. Processor cores maximised (-1) for parallel processing, accelerating model training. A random state, fixed at 777, persists, applying to all experimental procedures.

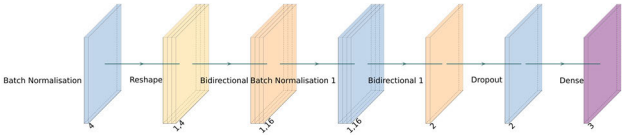


FIGURE 11. DL - Bidirectional recurrent neural network and Long-Short-Term memory neural network parameters (pump sensor).

TABLE 10. Deepchecks data explainability parameters parameters (pump sensor).

Parameter	Value	Description
n_samples	len(dataset)//2	The number of samples to use in the data integrity check
timeout	600	The maximum amount of time (in seconds) to allow for the check
n_top_columns	50	The number of top columns to include in the check
n_to_show	50	The number of results to show

2) PUMP SENSOR DATASET CONFIGURATIONS

The dataset was split, dedicating 70% for training and the rest for testing, ensuring uniformity in all trials with a random state of 777. It involved 4 distinct features as inputs and 3 classes as output for classification.

a: DEEP LEARNING (PUMP SENSOR)

In Fig. 11, the Bidirectional Recurrent Neural Network (Bi-RNN) and Long-Short-Term Memory Neural (LSTM) Network parameters were meticulously set. The input data are first transformed, normalised, and scaled. The data are then reshaped into a (1, 4) configuration, followed by a Bi-RNN and LSTM with 16 output channels. To enhance regularisation, a certain percentage of inputs are randomly dropped. The final step involves a dense layer with 3 output units, facilitating the classification into 3 classes.

i) DATA EXPLAINABILITY (PUMP SENSOR)

Table 10 shows the parameters for the Deepchecks data explanations.

ii) MODEL EXPLAINABILITY (PUMP SENSOR)

Table 11 shows the parameters for the DL SHAP model explanations.

iii) OUTCOME EXPLAINABILITY (PUMP SENSOR)

Table 12 shows the parameters for the explanations of the DL SHAP outcomes.

iv) END-USER EXPLAINABILITY (PUMP SENSOR)

Table 13 shows the parameters for the DL DiCE end-user explanations.

b: MACHINE LEARNING (PUMP SENSOR)

The parameter configurations for the Adaptive Boosting classifier are shown in (Fig. 12). The learning rate is set to 0.2, dictating the step size for model updates. The Adaptive

TABLE 11. DL - SHAP Model explainability parameters (pump sensor).

Parameter	Value	Description
sample_size	25	Number of samples taken from the training data for SHAP analysis
X_sub	df_train	Subset of the training data (excluding the target variable) used for SHAP analysis
explainer	SHAP explainer object	An object that can calculate SHAP values, created using the model's prediction function and sampled data
shap_values_multiple	df_test[0:sample_size]	The SHAP values for the test data
class_indices	All five classes	List of class indices for which SHAP summary plots will be generated
class_label	[‘comp1’, ‘comp2’, ‘comp3’, ‘comp4’, ‘none’]	The label of the class corresponding to the current class index

TABLE 12. DL - SHAP Outcome explainability parameters (pump sensor).

Parameter	Value	Description
sample_size	25	Number of samples taken from the training data for SHAP analysis
X_sub	df_train	Subset of the training data (excluding the target variable) used for SHAP analysis
explainer	SHAP explainer object	An object that can calculate SHAP values, created using the model's prediction function and sampled data
shap_values	df_test[0:1]	The SHAP values for the test data
class_indices	All five classes	List of class indices for which SHAP summary plots will be generated
class_label	[‘comp1’, ‘comp2’, ‘comp3’, ‘comp4’, ‘none’]	The label of the class corresponding to the current class index

```
AdaBoostClassifier
AdaBoostClassifier(learning_rate=0.2, n_estimators=100, random_state=777)
```

FIGURE 12. ML - Adaptive boosting classifier parameters (pump sensor).

Boosting classifier includes 100 estimators (weak models) by definition. A random state, fixed at 777, persists, applying to all experimental procedures.

IV. RESULTS AND DISCUSSION

This section presents the results of XAI and PdM in SAF in distinct sub-themes. The research delved into prognostics and diagnostics for PdM. Prognostics predicts water pump conditions, whereas diagnostics identifies specific machine failure components. Both use DL and ML algorithms. The study also explored the ‘‘data dimension’’, assessing how data limitations and expectations influence insight extraction. Moreover, this section presents global explainability, aiming to unravel

TABLE 13. DL - DiCE end-user explainability parameters (pump sensor).

Parameter	Value	Description
data_object	dice_ml.Data object	DiCE data object prepared using the input dataset
backend	'TF'+tf.__version__[0]	TensorFlow backend version used (e.g., 'TF2' for TensorFlow version 2.x)
model_object	dice_ml.Model object	DiCE model object created using the TensorFlow model
explainer	dice_ml.Dice object	DiCE explanation object instantiated with the data object, model object, and the method set to 'random'
desired_classes	All other classes	List of desired classes for which counterfactuals are generated
test_query	DataFrame object	Subset of a dataset (excluding the target variable) representing a test query for which counterfactuals are generated
total_CFs	5	Total number of counterfactual instances to generate for each desired class
features_to_vary	"all"	Specification of features to vary during counterfactual generation
proximity_weight	1.5	Weight is assigned to proximity in the counterfactual generation process. Feature-wise distance from the original input
diversity_weight	1.0	The weight assigned to diversity in the counterfactual generation process. Feature-wise distance between each counterfactual pair
stopping_threshold	0.5	Threshold for stopping the counterfactual generation process

the model's behaviour through model-centric approaches for the ''model dimension''. Local explainability strives to explain single prediction instances, providing insights for the ''outcome dimension''. Addressing the ''end-user dimension'' emphasises creating explanations that balance abstraction and detail. In addition, discussions detail a comparative analysis with related studies, while following themes (prognostics, diagnostics, data, model, outcome, and end-user) and results interpretation, exposing implications and alignment to the related literature; this section concludes by summarising results and discussion.

Recall, that this research had a dual aim to: (1) predict maintenance needs and (2) provide explanations using XAI in PdM for SAF. Section II conducted an SLR with bibliometric analysis, tackling the first research objective: (a) conducting an SLR using a bibliometric analysis to determine the current landscape of AI models and the need for XAI. Section III addressed the following three objectives of this research to achieve its aim: (b) developing an AI model to predict maintenance needs, (c) evaluating the proposed AI model, (d) then identifying and providing explanations for the predictions made by the proposed AI model.

A. EXPERIMENTAL RESULTS

The experimental results reveal insights from the PdM alongside the XAI experiments. First, it dissects the results of both the employed DL and ML models. Next, it shifts towards XAI, casting light on the four explainability dimensions drawn from scholarly literature: (1) data, (2) model, (3) outcome, and (4) end-user.

1) PREDICTIVE MAINTENANCE RESULTS

In Section II, the research navigated prognostics and diagnostics for PdM in SAF, focusing on system failure and component failure prediction literature. The Telemetry for Predictive Maintenance dataset, initially with 11 attributes, expanded to 40 through feature engineering and refined to 19 for optimised performance, dismissing non-essential attributes like ''Datatime'', ''MachineID'', and ''model''. Similarly, the Pump Sensor dataset, starting with 54 attributes, was distilled to 214, then to 4 after releasing irrelevant or incomplete attributes. With high test F1 scores (above 92%), accuracies (above 90%), and ROC-AUCs (above 80%), the models exhibited strong adaptability to new data, avoiding overfitting. DT and CatBoost classifiers, while slightly trailing, still delivered competitive results. These results, detailed in Tables 14 to 17, underscore the proposed model's efficacy as a leading solution for PdM in SAF.

a: DEEP LEARNING (PREDICTIVE MAINTENANCE RESULTS)

Table 14 shows the DL classifier comparison on the Telemetry for Predictive Maintenance dataset.

TABLE 14. DL - classifier comparison (telemetry for predictive maintenance).

Classifier	Train F1 score	Test F1 score	Overfitting	Train accuracy	Test accuracy	Test ROC-AUC
Convolutional LSTM Neural Net	0.958015	**0.970003**	False	0.942831	**0.942831**	0.967616
BiRNN LSTM Neural Net	0.95516	0.963522	False	0.945424	0.945424	0.957594
Artificial Neural Net	0.945871	0.95583	False	0.922676	0.922676	0.939171
Convolutional Neural Net	0.936568	0.939713	False	0.905862	0.905862	**0.908591**

Table 15 shows the DL classifier comparison on the Pump Sensor dataset.

TABLE 15. DL - classifier comparison (pump sensor).

Classifier	Train F1 score	Test F1 score	Overfitting	Train accuracy	Test accuracy	Test ROC-AUC
BiRNN LSTM Neural Net	0.862548	**0.998576**	False	0.90576	0.998782	0.928758
Artificial Neural Net	0.861264	0.998264	False	0.905963	**0.998843**	0.802616
Convolutional Neural Net	0.948287	0.997471	False	0.924936	0.996573	0.800625
Convolutional LSTM Neural Net	0.955991	0.960525	False	0.933075	0.92518	**0.935467**

b: MACHINE LEARNING (PREDICTIVE MAINTENANCE RESULTS)

Table 16 shows the ML classifier comparison on the Telemetry for Predictive Maintenance dataset.

TABLE 16. ML - classifier comparison (telemetry for predictive maintenance).

Classifier	Train F1 score	Test F1 score	Overfitting	Train accuracy	Test accuracy	Test ROC-AUC
Bagging classifier	0.993066	**0.994729**	False	0.992466	**0.992466**	0.994373
LGBM classifier	0.993131	0.99447	False	0.992275	0.992275	0.993802
Decision Tree classifier	0.993036	0.994258	False	0.992432	0.992432	0.993836
XGB classifier	0.993043	0.993304	False	0.993733	0.993733	0.993779
RandomForest classifier	0.985309	0.985562	False	0.982755	0.982755	0.982501
CatBoost classifier	0.983746	0.984394	False	0.98028	0.98028	**0.980435**
AdaBoost classifier	0.969947	0.97304	False	0.977849	0.977849	0.981074

Table 17 shows the ML classifier comparison on the Pump Sensor dataset.

Furthermore, as established earlier, applying the test F1 score calls for its use in selecting distinguished DL and ML algorithms per dataset [32], [44]. Thus, the Convolutional LSTM Neural Network, BiRNN LSTM Neural Network, Bagging classifier, and AdaBoost classifier algorithms were used for the explainability stage.

TABLE 17. ML - classifier comparison (pump sensor).

Classifier	Train F1 score	Test F1 score	Overfitting	Train accuracy	Test accuracy	Test ROC-AUC
AdaBoost classifier	0.945365	**0.999525**	False	0.951154	**0.999574**	**0.969477**
RandomForest classifier	0.99794	0.998902	False	0.997931	0.998766	0.936051
LGBM classifier	0.996472	0.998612	False	0.996417	0.998919	0.905535
XGB classifier	0.995269	0.998347	False	0.995255	0.998051	0.906245
Decision Tree classifier	0.971186	0.992829	False	0.969703	0.98803	0.494588
Bagging classifier	0.971175	0.992821	False	0.96969	0.988015	0.832808
CatBoost classifier	0.886011	0.951349	False	0.838352	0.90922	0.492795

2) EXPLAINABLE ARTIFICIAL INTELLIGENCE RESULTS

Section III's "Data" segment, showed two datasets: Telemetry for Predictive Maintenance and Pump Sensor, each with 5 and 3 classes, respectively. The analysis considers these class structures, unless otherwise indicated. Key variables like "age", "datetime_day_of_month", "datetime_hour_sin", and "datetime_week" transformed into float values post-standardisation; this section presents and discusses explainability dimensions, using DL algorithms as input tools.

a: INSIGHTS FROM THE DATA

Using data-centric methods, the study extracted deep insights directly from the data, assessing its purity and quality. Results showed meticulous dataset preparation, adhering to consistency, integrity, and predictability criteria. For instance, the Telemetry for Predictive Maintenance dataset featured 21 columns, surpassing the original 19. In contrast, the Pump Sensor dataset had 6, exceeding the original 4. However, these extra columns, included categorical features and target variables, which were essential metadata, influencing neither data purity checks nor overall test outcomes. Data evaluations in Table 18 showcased successful checks, confirming diverse values, the absence of excessive special characters, uniform handling of null values, minimal variation in data types, string content uniformity, lack of duplicate data, and relevant correlations under the threshold. A minor discrepancy emerged, with 0.02% of the samples showing conflicting labels, narrowly exceeding the set limit. Similarly, Table 19 affirmed diversity in columns, appropriate handling of special characters and null values, minimal data-type variation, string uniformity, and no duplicate data. Although, feature-feature correlations and predictive power of features like "sensor_05", with a score of 0.93, stood out, a single test narrowly missed the mark. These evaluations showed the

TABLE 18. Data purity (telemetry for predictive maintenance).

Status	Check	Condition	More Information
Passed	Single Value in Column	It does not contain only a single value	Passed for 21 relevant columns
Passed	Special Characters	The ratio of samples containing special characters is less or equal to 0.1%	Passed for 21 relevant columns
Passed	Mixed Nulls	The number of different null types is less or equal to 1	Passed for 21 relevant columns
Passed	Mixed Data Types	Rare data types in column are either more than 10% or less than 1% of the data	21 columns passed: found 0 columns with negligible types mix, and 21 columns without any types mix
Passed	String Mismatch	No string variants	Passed for one relevant column
Passed	Data Duplicates	Duplicate data ratio is less or equal to 5%	Found 0% duplicate data
Passed	String Length Out of Bounds	The ratio of string length outliers is less or equal to 0%	No relevant columns to check were found
Passed	Feature-Label Correlation	Features' Predictive Power Score (PPS) is less than 0.8	Passed for 19 relevant columns
Passed	Feature-Feature Correlation	Not more than 0 pairs are correlated above 0.9	All correlations are less than 0.9 except for pairs
Failed	Conflicting Labels	Ambiguous sample ratio is less or equal to 0%	The ratio of samples with conflicting labels: 0.02%

TABLE 19. Data purity (pump sensor).

Status	Check	Condition	More Information
Passed	Single Value in Column	It does not contain only a single value	Passed for six relevant columns
Passed	Special Characters	The ratio of samples containing special characters is less or equal to 0.1%	Passed for six relevant columns
Passed	Mixed Nulls	The number of different null types is less or equal to 1	Passed for six relevant columns
Passed	Mixed Data Types	Rare data types in column are either more than 10% or less than 1% of the data	Six columns passed: found 0 columns with negligible types mix, and six columns without any types mix
Passed	String Mismatch	No string variants	Passed for one relevant column
Passed	Data Duplicates	Duplicate data ratio is less or equal to 5%	Found 0% duplicate data
Passed	String Length Out of Bounds	The ratio of string length outliers is less or equal to 0%	No relevant columns to check were found
Passed	Conflicting Labels	Ambiguous sample ratio is less or equal to 0%	The ratio of samples with conflicting labels: 0%
Passed	Feature-Feature Correlation	Not more than 0 pairs are correlated above 0.9	All correlations are less than 0.9 except for pairs
Failed	Feature-Label Correlation	Features' Predictive Power Score is less than 0.8	Found 3 out of 4 features with PPS above threshold: sensor_05: 0.93 sensor_10_window_3H_mean: 0.83 sensor_12_window_3H_mean: 0.81

passing of 9 out of 10 tests in both datasets, underscoring the datasets' robustness and reliability, setting a strong foundation for subsequent PdM and AI analysis.

i) DATA RESULTS

Table 18 shows the data purity tests on the Telemetry for Predictive Maintenance dataset. Table 19 shows the data purity tests on the Pump Sensor dataset.

b: INSIGHTS FROM THE MODEL

A focus on global explainability emerged, and delved into model-centric methodologies that explore links between input features and outcomes; this approach yielded rich insights directly from the model, primarily through SHAP plots, which clarify feature impact direction using colour-coded bars (blue for negative, red for positive). For clarity, plot titles like "class 'compX'" indicate "compY's" influence on the target "compX", underpinning precise component-level PdM. Unlike feature importance, SHAP values provide a nuanced view of feature attribution, applying game theory for a comprehensive impact assessment [68], [69]. Also, unlike feature importance, which ranks features by performance but lacks detailed impact analysis [70], SHAP values offer a balanced view, suitable for both global and local effects. The research also leverages SHAP for global feature attribution using DL algorithms, providing a holistic understanding of model behaviour. For instance, Fig. 13 shows how the "error1count_1" feature significantly influences predictions for class "comp1," with a notable SHAP value of 0.09. Similarly, Fig. 14 highlights sensor_05 as a key predictor for a "broken" pump status, shown by its leading SHAP value; this analysis, extending beyond individual predictions, helps identify critical features for effective PdM, guiding future data strategies.

i) DEEP LEARNING (MODEL RESULTS)

Fig. 13 shows the DL Global SHAP for class comp1 on the Telemetry for Predictive Maintenance dataset.

Fig. 14 shows the DL Global SHAP for class broken on the Pump Sensor dataset.

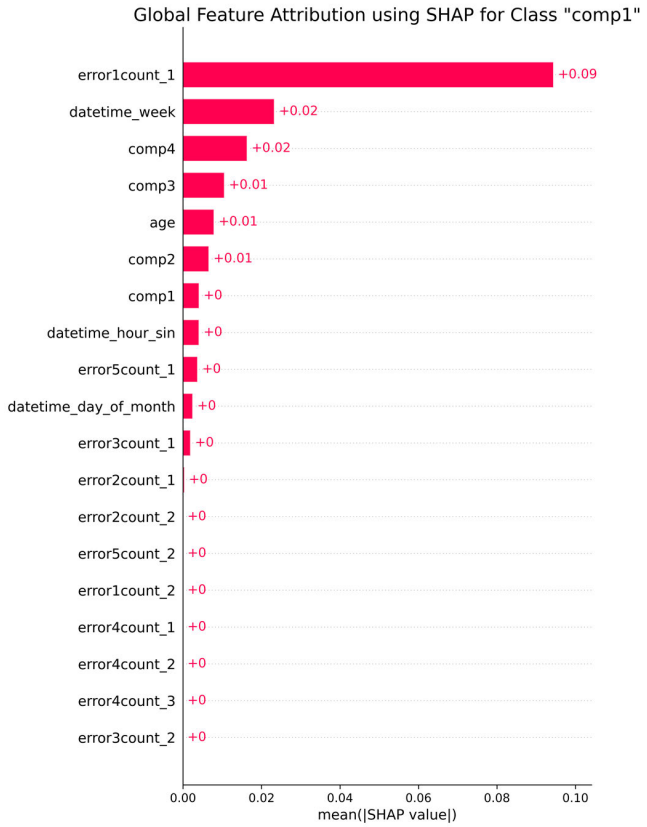


FIGURE 13. DL - Global SHAP for class comp1 (telemetry for predictive maintenance).

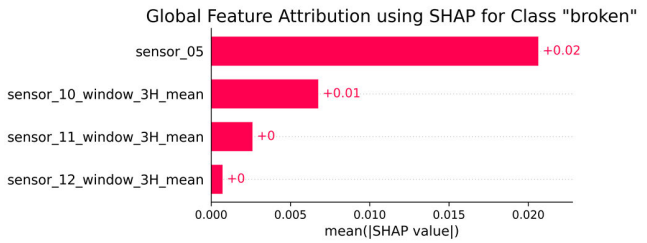


FIGURE 14. DL - Global SHAP for class broken (pump sensor).

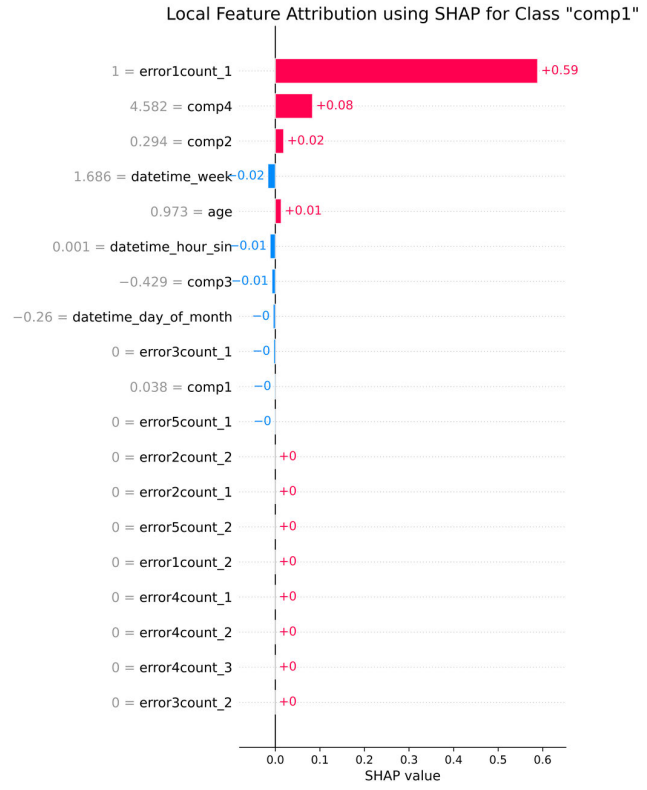


FIGURE 15. DL - Local SHAP for class comp1 (telemetry for predictive maintenance).

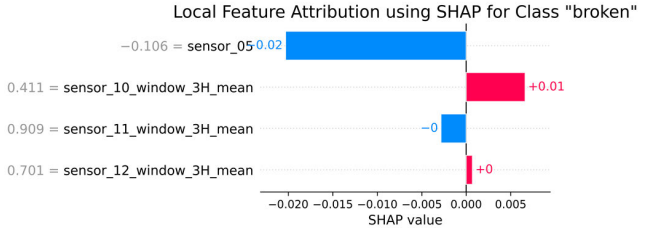


FIGURE 16. DL - Local SHAP for class broken (pump sensor).

c: INSIGHTS FROM THE OUTCOME

Local explainability methods illuminate individual prediction instances, extending the model's insights. SHAP, central to this analysis, analyses single-instance feature impact. Features like "error1count_2" quantify specific error occurrences, while "datetime_hour_sin" and "datetime_hour_cos" emerge from sine and cosine transformations (detailed in Section III's "Periodic Cyclic Features"). The study then applied SHAP to DL for local feature attribution in the outcome dimension, providing detailed insights. Unlike global attribution, local SHAP provides granular explainability. For instance, in Fig. 15, "error1count_1" significantly sways the prediction towards "comp1". Other features like "comp4" and "comp2" also impact this prediction, albeit less so. On the contrary, "datetime_week", "datetime_hour_sin", and "comp3" reduce the impact of classifying as "comp1". Features such as "error3count_1" and "comp1" show zero impact on classification, highlighting the nuanced influence of different features. Fig. 16 further shows local attribution, where "sensor_05" negatively affects the "broken" pump status prediction, while "sensor_10_window_3H_mean" enhances it. However, features like "sensor_11_window_3H_mean" and "sensor_12_window_3H_mean" show no effect on the prediction; this approach underscores the specific influence of each feature on the model's decision-making process.

i) DEEP LEARNING (OUTCOME RESULTS)

Fig. 15 shows the DL Local SHAP for class comp1 on the Telemetry for Predictive Maintenance dataset.

Fig. 16 shows the DL Local SHAP for class broken on the Pump Sensor dataset.

d: INSIGHTS TOWARD THE END-USER

CFEs balance abstraction and detail. ''Counterfactual n for class 'compX'" indicates altering ''compY'' to trigger a ''compX'' a broken status, aiding in pinpointing the malfunctioning component. Phrases like ''The feature compY must transition'' indicate necessary changes in ''compY'' to affect ''compX''. Some CFEs, without percentage change, relate to categorical features, while others, exceeding 100%, indicate the extent of change needed for a desired outcome. Light grey and light blue in the CFE tables represent decreasing and increasing feature changes, respectively. Table 20 presents ''what-if'' (counterfactuals) scenarios for the Telemetry for Predictive Maintenance dataset, showing changes needed to switch from class ''none'' to ''comp1''. These counterfactuals highlight the impact of features like ''comp4'' and ''comp2'', with substantial percentage changes (increases of 593.70% and decreases of 165.64%, respectively). The ''errorXcount_Y'' features, moving from 0.0 to 1.0 demonstrate the influence of specific errors. Counterfactuals reveal that minor adjustments, like a 22.33% change in ''datetime_week'', can be significant. Consistent settings of ''error1count_2'' and ''error2count_1'' to 1.0 in Counterfactual five highlight their ongoing effect on outcomes. In the Pump Sensor dataset, counterfactuals in Table 21 show shifts from ''normal'' to ''broken''. A dramatic change in ''sensor_05'' is required, indicating its crucial role in predicting breakdowns. Counterfactuals two and three also demand significant adjustments in ''sensor_12_window_3H_mean'' and ''sensor_05''. ''Sensor_05'' is consistently altered across scenarios, underscoring its impact on classifying the ''broken'' status.

i) DEEP LEARNING (END-USER RESULTS)

Table 20 shows the DL CFE from none to comp1 on the Telemetry for Predictive Maintenance dataset.

TABLE 20. DL - CFE from none to comp1 (telemetry for predictive maintenance).

Counterfactual one for class "comp1"
The feature comp4 must transition from -0.369327187538147 to 1.82338613 **(593.70%)**
The feature error4count_2 must transition from 0.0 to 1.0
Counterfactual two for class "comp1"
The feature comp1 must transition from 2.1010522842407227 to 4.61362886 **(119.59%)**
The feature error5count_2 must transition from 0.0 to 1.0
Counterfactual three for class "comp1"
The feature datetime_week must transition from -1.7300148010253906 to -1.3437247 **(22.33%)**
The feature error5count_2 must transition from 0.0 to 1.0
Counterfactual four for class "comp1"
The feature comp2 must transition from 1.2788968086242676 to -0.83951104 **(-165.64%)**
The feature error1count_1 must transition from 0.0 to 1.0
Counterfactual five for class "comp1"
The feature error1count_2 must transition from 0.0 to 1.0
The feature error2count_1 must transition from 0.0 to 1.0

Table 21 shows the DL CFE from normal to broken on the Pump Sensor dataset.

TABLE 21. DL - CFE from normal to broken (pump sensor).

Counterfactual one for class "broken"
The feature sensor_05 must transition from 0.1950564682483673 to -3.5593897 **(-1924.80%)**
Counterfactual two for class "broken"
The feature sensor_12_window_3H_mean must transition from 0.7513946890830994 to -2.0743027 **(-376.06%)**
The feature sensor_05 must transition from 0.1950564682483673 to -3.0883916 **(-1683.33%)**
Counterfactual three for class "broken"
The feature sensor_12_window_3H_mean must transition from 0.7513946890830994 to -2.2550662 **(-400.12%)**
The feature sensor_05 must transition from 0.1950564682483673 to -3.3266737 **(-1805.49%)**
Counterfactual four for class "broken"
The feature sensor_05 must transition from 0.1950564682483673 to -3.639838 **(-1966.04%)**
Counterfactual five for class "broken"
The feature sensor_05 must transition from 0.1950564682483673 to -3.523411 **(-1906.35%)**

B. DISCUSSION

After presenting the experimental results, the results are then discussed in the context of this research. It starts with a comparative analysis with related studies while discussing the results in a similar order as was done in the presentation of results: PdM and the XAI results. The first (PdM) discusses the results of both the DL and ML models. The second (XAI) discusses the results of XAI, focusing on the four dimensions of explainability extracted from the literature: (1) data, (2) model, (3) outcome, and (4) end-user.

1) COMPARATIVE ANALYSIS WITH RELATED STUDIES

The comparative analysis in this section offers insights into this research's unique contributions, contrasting with previous studies to deepen understanding. Performance comparisons in Table 22 demonstrate the superior functioning of DL and ML classifiers on two datasets. These models, Convolutional LSTM and BiRNN LSTM Neural Nets, excel across metrics without overfitting, highlighting their reliable generalisation. Comparisons with [33] show that this research's LSTM classifiers consistently outperformed others, with a 5.81% test accuracy improvement, calculated as an average percentage difference. Moreover, the XGBoost classifiers perform exceptionally well on diverse datasets (Table 23). For instance, in the Telemetry for Predictive Maintenance dataset, XGBoost achieves consistent F1 scores and accuracy. Compared to [44], this research's XGBoost classifier shows enhanced performance, with an increase of 7.09% in test F1 score, 10.66% in accuracy, and 4.29% in ROC-AUC;

this research's XGBoost classifier surpasses [44] in all test metrics, affirming the superiority of this approach in PdM, machine status, and water pump data analysis. These results emphasise LSTMs and Boosters' efficacy in PdM for SAF.

Table 22 shows the DL classifier comparison with related studies.

TABLE 22. DL - classifier comparison with related studies.

Source	Dataset	Classifier	Train F1 score	Test F1 score	Overfitting	Train accuracy	Test accuracy	Test ROC-AUC
This research	Telemetry for Predictive Maintenance	Convolutional LSTM Neural Net	0.958 015	0.970 003	False	0.942 831	0.942 831	**0.967 616**
This research	Telemetry for Predictive Maintenance	BiRNN LSTM Neural Net	0.955 16	0.963 522	False	0.945 424	0.945 424	0.957 594
This research	Pump Sensor	BiRNN LSTM Neural Net	0.862 548	**0.998 576**	False	0.905 76	**0.998 782**	0.928 758
This research	Pump Sensor	Convolutional LSTM Neural Net	0.955 991	0.960 525	False	0.933 075	0.925 18	0.935 467
[33]	Time series	RNN LSTM	-	-	-	-	0.900 7	-

Table 23 shows the ML classifier comparison with related studies.

TABLE 23. ML - classifier comparison with related studies.

Source	Dataset	Classifier	Train F1 score	Test F1 score	Overfitting	Train accuracy	Test accuracy	Test ROC-AUC
This research	Telemetry for Predictive Maintenance	XGB classifier	0.993 043	0.993 304	False	0.993 733	0.993 733	**0.993 779**
This research	Pump Sensor	XGB classifier	0.995 269	**0.998 347**	False	0.995 255	**0.998 051**	0.906 245
[44]	Time series	XGB classifier	-	0.93	-	-	0.9	0.911

2) EXPLAINABILITY DIMENSIONS

a: DATA DIMENSION

Contrary to related studies, this research addresses an unexplored dimension: data explainability. Pioneering in this dimension, the study sets a precedent for future research. It methodically demonstrates how to discern data limitations and align researcher expectations. Tables 18 and 19 exhibit rigorous dataset analysis, highlighting diversity, uniformity, and minimal duplications. Despite minor label discrepancies, these datasets exemplify consistency and integrity, crucial for reliable PdM and AI analysis.

b: MODEL AND OUTCOME DIMENSIONS

This research diverges from related studies in its model and outcome explainability approach. A key premise often assumed is the independence and non-correlation of features or attributes; this assumption is fallible, introducing potential biases in the ''model and outcome dimensions.'' Thus, data explainability becomes a foundational element for these dimensions, guiding the assessment of inherent limitations and depth of insights derivable from the data; this research's contribution lies in establishing a framework for exploring the levels of explainability achievable with specific datasets. Employing SHAP for global feature attribution in DL, it transcends mere feature ranking, enabling a comprehensive understanding of model behaviour. In local explainability, the study goes further, analysing individual prediction instances. Local methods, through SHAP, provide an in-depth feature impact analysis on single-instance predictions.

c: END-USER DIMENSION

Concerning end-user explainability, this research stands out from existing literature, particularly in explaining predictions to non-technical users; this study significantly advances end-user explainability, showcasing how to strike this balance effectively. Using the two datasets, it presented ''what-if'' (counterfactuals) scenarios in Tables 20 and 21, demonstrating the necessary changes to shift predictions from one class to another preferred class (outcome). These counterfactuals elucidate the impact of specific features, underscoring the significance of consistent settings in influencing outcomes.

3) SCOPE AND LIMITATIONS OF THE RESEARCH

The scope of the study was to incorporate XAI and PdM in SAF, focusing on (1) machines and (2) irrigation systems. Moreover, the study faced constraints stemming from limited PdM data, using merely two time series datasets. PdM findings focused exclusively on DL and EL for ML models, influenced by pertinent literature. Similarly, the scope of XAI was confined to (1) data, (2) model, (3) outcome, and (4) end-user explanations, framed by established dimensions of explainability. Additionally, the bibliometric analysis faced its own set of limitations; the review methodology encapsulated data processes, descriptive statistics, keyword analysis, knowledge synthesis, and the exploration of conceptual and social structures. Searches targeted ''Titles'', ''Abstracts'', and ''Subject Headings'' in a period from 2012 to 2022, excluding the incomplete year 2023. In the databases, ''Keywords Plus'' and ''Keywords'' were used as Subject Headings to search for synonyms, domain-specific language, and additional known relevant words [49]; this approach extracted literature relevant to the topic of this research. In addition, only two academic databases were used: WoS and Scopus. For replication and ensuring reproducibility, the researcher's experimental code is available in their GitHub repository at: https://github.com/iammelvink.

V. CONCLUSION

This research marks a significant advancement in SAF, merging XAI with PdM. It predicts maintenance requirements and elucidates model predictions; this fusion enhances understanding and application of AI for PdM in SAF, addressing

the research aim and objectives. The study illuminates the potent synergy between XAI and PdM, fostering transparent, understandable algorithmic decisions, a vital move towards clarifying the often opaque nature of DL and ML. The research showcases the impressive performance of the LSTM and XGBoost classifiers in PdM, setting new benchmarks in SAF. These advances promise improved predictive accuracy and reliability. Exploring data dimensions reveals inherent limitations and refines expectations for data-driven insights. The research unravels the model and outcome complexities using SHAP values and providing detailed, practical explanations. It also introduces varied predictive scenarios through counterfactuals, which could improve stakeholder decision-making.

Future research directions include integrating multi-modal data to enhance AI model efficacy for PdM in SAF; developing new explainability metrics; incorporating Human-in-the-Loop (HITL) systems for collaborative decision-making and enhanced PdM model reliability; analysing the effects of XAI on human decision-making in agriculture; exploring ethical aspects of agricultural AI, focusing on Fairness, Accountability, and Transparency (FAT); and assessing the long-term impact of these technologies on productivity, sustainability, and economic factors in SAF, setting the stage for future studies that are as impactful as they are essential.

REFERENCES

[1] L. W. Bell, A. D. Moore, and J. A. Kirkegaard, "Evolution in crop–livestock integration systems that improve farm productivity and environmental performance in Australia," (in English), *Eur. J. Agronomy*, vol. 57, pp. 10–20, Jul. 2014, doi: 10.1016/j.eja.2013.04.007.

[2] J. Rana and J. Paul, "Consumer behavior and purchase intention for organic food: A review and research agenda," (in English), *J. Retailing Consum. Services*, vol. 38, pp. 157–165, Sep. 2017, doi: 10.1016/j.jretconser.2017.06.004.

[3] Y. Zhong, I. K. W. Lai, F. Guo, and H. Tang, "Research on government subsidy strategies for the development of agricultural products e-commerce," (in English), *Agriculture*, vol. 11, no. 11, p. 1152, Nov. 2021, doi: 10.3390/agriculture11111152.

[4] A. Calcante, L. Fontanini, and F. Mazzetto, "Repair and maintenance costs of 4WD tractors in northern Italy," (in English), *Trans. ASABE*, vol. 56, no. 2, pp. 355–362, 2013, doi: 10.13031/2013.42660.

[5] E. Elahi, Z. Khalid, M. Z. Tauni, H. Zhang, and X. Lirong, "Extreme weather events risk to crop-production and the adaptation of innovative management strategies to mitigate the risk: A retrospective survey of rural Punjab, Pakistan," (in English), *Technovation*, vol. 117, Sep. 2022, Art. no. 102255, doi: 10.1016/j.technovation.2021.102255.

[6] M. Yildirim, N. Z. Gebraeel, and X. A. Sun, "Integrated predictive analytics and optimization for opportunistic maintenance and operations in wind farms," (in English), *IEEE Trans. Power Syst.*, vol. 32, no. 6, pp. 4319–4328, Nov. 2017, doi: 10.1109/TPWRS.2017.2666722.

[7] P. Zhou and P. T. Yin, "An opportunistic condition-based maintenance strategy for offshore wind farm based on predictive analytics," (in English), *Renew. Sustain. Energy Rev.*, vol. 109, pp. 1–9, Jul. 2019, doi: 10.1016/j.rser.2019.03.049.

[8] C. Eastwood, L. Klerkx, M. Ayre, and B. Dela Rue, "Managing socio-ethical challenges in the development of smart farming: From a fragmented to a comprehensive approach for responsible research and innovation," (in English), *J. Agricult. Environ. Ethics*, vol. 32, nos. 5–6, pp. 741–768, Dec. 2019, doi: 10.1007/s10806-017-9704-5.

[9] S. Wolfert, L. Ge, C. Verdouw, and M.-J. Bogaardt, "Big data in smart farming—A review," (in English), *Agricult. Syst.*, vol. 153, pp. 69–80, May 2017, doi: 10.1016/j.agsy.2017.01.023.

[10] S. A. Z. Rahman, K. C. Mitra, and S. M. M. Islam, "Soil classification using machine learning methods and crop suggestion based on soil series," in *Proc. 21st Int. Conf. Comput. Inf. Technol. (ICCIT)*, Dec. 2018, pp. 1–4. [Online]. Available: https://ieeexplore.ieee.org/document/8631943/

[11] V. Panchbhaiyye and T. Ogunfunmi, "Experimental results on using deep learning to identify agricultural pests," in *Proc. IEEE Global Humanitarian Technol. Conf. (GHTC)*, Oct. 2018, pp. 1–2. [Online]. Available: https://ieeexplore.ieee.org/document/8601896/

[12] P. Shankar, N. Werner, S. Selinger, and O. Janssen, "Artificial intelligence driven crop protection optimization for sustainable agriculture," in *Proc. IEEE/ITU Int. Conf. Artif. Intell. Good (AI4G)*, Sep. 2020, pp. 1–6. [Online]. Available: https://ieeexplore.ieee.org/document/9311082/

[13] N. Taravatrooy, M. R. Nikoo, S. Hobbi, M. Sadegh, and A. Izady, "A novel hybrid entropy-clustering approach for optimal placement of pressure sensors for leakage detection in water distribution systems under uncertainty," (in English), *Urban Water J.*, vol. 17, no. 3, pp. 185–198, Mar. 2020, doi: 10.1080/1573062x.2020.1758162.

[14] D. R. Vincent, N. Deepa, D. Elavarasan, K. Srinivasan, S. H. Chauhdary, and C. Iwendi, "Sensors driven AI-based agriculture recommendation model for assessing land suitability," *Sensors*, vol. 19, no. 17, p. 3667, Aug. 2019, doi: 10.3390/s19173667.

[15] M. Kande, A. Isaksson, R. Thottappillil, and N. Taylor, "Rotating electrical machine condition monitoring automation—A review," *Machines*, vol. 5, no. 4, p. 24, Oct. 2017, doi: 10.3390/machines5040024.

[16] K. Poppe, S. Wolfert, C. Verdouw, and A. Renwick, "A European perspective on the economics of big data," *Farm Policy J.*, vol. 12, no. 1, pp. 11–19, 2015.

[17] S. Sonka, "Big data: From hype to agricultural tool," *Farm Policy J.*, vol. 12, pp. 1–9, Jan. 2015. [Online]. Available: https://www.researchgate.net/publication/279771638_Big_Data_From_Hype_to_Agricultural_Tool

[18] A. Brauneck, L. Schmalhorst, M. M. K. Majdabadi, M. Bakhtiari, U. Völker, J. Baumbach, L. Baumbach, and G. Buchholtz, "Federated machine learning, privacy-enhancing technologies, and data protection laws in medical research: Scoping review," *J. Med. Internet Res.*, vol. 25, Mar. 2023, Art. no. e41588, doi: 10.2196/41588.

[19] T. R. Chhetri, A. Kurteva, R. J. DeLong, R. Hilscher, K. Korte, and A. Fensel, "Data protection by design tool for automated GDPR compliance verification based on semantically modeled informed consent," *Sensors*, vol. 22, no. 7, p. 2763, Apr. 2022, doi: 10.3390/s22072763.

[20] A. D. Selbst and S. Barocas, "The intuitive appeal of explainable machines," (in English), *Fordham Law Rev.*, vol. 87, no. 3, pp. 1085–1139, Dec. 2018.

[21] E. Lughofer and M. Sayed-Mouchaweh, *Predictive Maintenance in Dynamic Systems*. Cham, Switzerland: Springer, 2019, doi: 10.1007/978-3-030-05645-2.

[22] F. Doshi-Velez and B. Kim, "Towards a rigorous science of interpretable machine learning," 2017, *arXiv:1702.08608*.

[23] D. Orn, L. Duan, Y. Liang, H. Siy, and M. Subramaniam, "Agro-AI education: Artificial intelligence for future farmers," in *Proc. 21st Annu. Conf. Inf. Technol. Educ.*, Oct. 2020, pp. 54–57, doi: 10.1145/3368308.3415457.

[24] V. P. Harmani, B. M. Himawan, M. A. Alhadi, and A. A. S. Gunawan, "Systematic literature review: Implementation of artificial intelligence in precision agriculture," in *Proc. 5th Int. Conf. Inf. Commun. Technol. (ICOIACT)*, Aug. 2022, pp. 479–484. [Online]. Available: https://ieeexplore.ieee.org/document/9971917/

[25] U. Ehsan, P. Wintersberger, Q. V. Liao, M. Mara, M. Streit, S. Wachter, A. Riener, and M. O. Riedl, "Operationalizing human-centered perspectives in explainable AI," in *Proc. Extended Abstr. CHI Conf. Human Factors Comput. Syst.*, Yokohama, Japan, May 2021, pp. 1–6, doi: 10.1145/3411763.3441342.

[26] S. Hepenstal, L. Zhang, and B. L. William Wong, "An analysis of expertise in intelligence analysis to support the design of human-centered artificial intelligence," in *Proc. IEEE Int. Conf. Syst., Man, Cybern. (SMC)*, Oct. 2021, pp. 107–112. [Online]. Available: https://ieeexplore.ieee.org/document/9659095/

[27] M. O. Riedl, "Human-centered artificial intelligence and machine learning," *Hum. Behav. Emerg. Technol.*, vol. 1, no. 1, pp. 33–36, Jan. 2019, doi: 10.1002/hbe2.117.

[28] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. H. Chau, A. Endert, and D. Keim, "A survey of human-centered evaluations in human-centered machine learning," (in English), *Comput. Graph. Forum*, vol. 40, no. 3, pp. 543–568, Jun. 2021, doi: 10.1111/cgf.14329.

[29] C. Molnar, G. Casalicchio, and B. Bischl, "Interpretable machine learning—A brief history, state-of-the-art and challenges," in *Proc. ECML PKDD Workshops*. Berlin, Germany: Springer, 2020, ch. 28, pp. 417–431.

[30] E. Kononov, A. Klyuev, and M. Tashkinov, "Prediction of technical state of mechanical systems based on interpretive neural network model," (in English), *Sensors*, vol. 23, no. 4, p. 1892, Feb. 2023, doi: 10.3390/s23041892.

[31] B. Ghasemkhani, O. Aktas, and D. Birant, "Balanced K-star: An explainable machine learning method for Internet-of-Things-enabled predictive maintenance in manufacturing," (in English), *Machines*, vol. 11, no. 3, p. 322, Feb. 2023, doi: 10.3390/machines11030322.

[32] M. Gashi, B. Mutlu, and S. Thalmann, "Impact of interdependencies: Multi-component system perspective toward predictive maintenance based on machine learning and XAI," (in English), *Appl. Sci.*, vol. 13, no. 5, p. 3088, Feb. 2023, doi: 10.3390/app13053088.

[33] H. Wu, A. Huang, and J. W. Sutherland, "Layer-wise relevance propagation for interpreting LSTM-RNN decisions in predictive maintenance," (in English), *Int. J. Adv. Manuf. Technol.*, vol. 118, nos. 3–4, pp. 963–978, Jan. 2022, doi: 10.1007/s00170-021-07911-9.

[34] J. Jakubowski, P. Stanisz, S. Bobek, and G. J. Nalepa, "Anomaly detection in asset degradation process using variational autoencoder and explanations," (in English), *Sensors*, vol. 22, no. 1, p. 291, Dec. 2021, doi: 10.3390/s22010291.

[35] G. Hajgató, R. Wéber, B. Szilágyi, B. Tóthpál, B. Gyires-Tóth, and C. Hős, "PredMaX: Predictive maintenance with explainable deep convolutional autoencoders," (in English), *Adv. Eng. Informat.*, vol. 54, Oct. 2022, Art. no. 101778, doi: 10.1016/j.aei.2022.101778.

[36] M. Garouani, A. Ahmad, M. Bouneffa, M. Hamlich, G. Bourguin, and A. Lewandowski, "Towards big industrial data mining through explainable automated machine learning," (in English), *Int. J. Adv. Manuf. Technol.*, vol. 120, nos. 1–2, pp. 1169–1188, May 2022, doi: 10.1007/s00170-022-08761-9.

[37] S. J. Upasane, H. Hagras, M. H. Anisi, S. Savill, I. Taylor, and K. Manousakis, "A big bang-big crunch type-2 fuzzy logic system for explainable predictive maintenance," in *Proc. IEEE Int. Conf. Fuzzy Syst. (FUZZ-IEEE)*, Jul. 2021, pp. 1–8. [Online]. Available: https://ieeexplore.ieee.org/document/9494540/

[38] M. Hermansa, M. Kozielski, M. Michalak, K. Szczyrba, L. Wróbel, and M. Sikora, "Sensor-based predictive maintenance with reduction of false alarms—A case study in heavy industry," (in English), *Sensors*, vol. 22, no. 1, p. 226, Dec. 2021, doi: 10.3390/s22010226.

[39] M. Kuzlu, U. Cali, V. Sharma, and Ö. Güler, "Gaining insight into solar photovoltaic power generation forecasting utilizing explainable artificial intelligence tools," (in English), *IEEE Access*, vol. 8, pp. 187814–187823, 2020, doi: 10.1109/ACCESS.2020.3031477.

[40] O. Serradilla, E. Zugasti, J. Ramirez de Okariz, J. Rodriguez, and U. Zurutuza, "Adaptable and explainable predictive maintenance: Semi-supervised deep learning for anomaly detection and diagnosis in press machine data," (in English), *Appl. Sci.*, vol. 11, no. 16, p. 7376, Aug. 2021, doi: 10.3390/app11167376.

[41] O. Mey and D. Neufeld, "Explainable AI algorithms for vibration data-based fault detection: Use case-adadpted methods and critical evaluation," (in English), *Sensors*, vol. 22, no. 23, p. 9037, Nov. 2022, doi: 10.3390/s22239037.

[42] H. Choi, D. Kim, J. Kim, J. Kim, and P. Kang, "Explainable anomaly detection framework for predictive maintenance in manufacturing systems," (in English), *Appl. Soft Comput.*, vol. 125, Aug. 2022, Art. no. 109147, doi: 10.1016/j.asoc.2022.109147.

[43] R. Langone, A. Cuzzocrea, and N. Skantzos, "Interpretable anomaly prediction: Predicting anomalous behavior in Industry 4.0 settings via regularized logistic regression tools," (in English), *Data Knowl. Eng.*, vol. 130, Nov. 2020, Art. no. 101850, doi: 10.1016/j.datak.2020.101850.

[44] B. Steurtewagen and D. Van den Poel, "Adding interpretability to predictive maintenance by machine learning on sensor data," (in English), *Comput. Chem. Eng.*, vol. 152, Sep. 2021, Art. no. 107381, doi: 10.1016/j.compchemeng.2021.107381.

[45] M. Reif, F. Shafait, M. Goldstein, T. Breuel, and A. Dengel, "Automatic classifier selection for non-experts," (in English), *Pattern Anal. Appl.*, vol. 17, no. 1, pp. 83–96, Feb. 2014, doi: 10.1007/s10044-012-0280-z.

[46] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," (in English), *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, Dec. 2006, doi: 10.1016/j.neucom.2005.12.126.

[47] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020, doi: 10.1016/j.inffus.2019.12.012.

[48] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, D. Pedreschi, and F. Giannotti, "A survey of methods for explaining black box models," 2018, *arXiv:1802.01933*.

[49] E. Garfield, "KeyWords plus-ISI's breakthrough retrieval method. 1. Expanding your searching power on current-contents on diskette," *Curr Contents*, vol. 32, pp. 5–9, Aug. 1990.

[50] M. J. Page, J. E. McKenzie, P. M. Bossuyt, I. Boutron, T. C. Hoffmann, C. D. Mulrow, L. Shamseer, J. M. Tetzlaff, E. A. Akl, S. E. Brennan, and R. Chou, "The PRISMA 2020 statement: An updated guideline for reporting systematic reviews," *Systematic Rev.*, vol. 10, no. 1, p. 89, Dec. 2021, doi: 10.1186/s13643-021-01626-4.

[51] M. Aria and C. Cuccurullo, "Bibliometrix: An R-tool for comprehensive science mapping analysis," *J. Informetrics*, vol. 11, no. 4, pp. 959–975, Nov. 2017, doi: 10.1016/j.joi.2017.08.007.

[52] J. Luengo, D. García-Gil, S. Ramírez-Gallego, S. García, and F. Herrera, *Big Data Preprocessing*. Cham, Switzerland: Springer, 2020, pp. 101–119.

[53] *Posit*. Accessed: Jun. 28, 2023. [Online]. Available: https://posit.co/downloads

[54] M. J. Cobo, A. G. López-Herrera, E. Herrera-Viedma, and F. Herrera, "Science mapping software tools: Review, analysis, and cooperative study among tools," (in English), *J. Amer. Soc. Inf. Sci. Technol.*, vol. 62, no. 7, pp. 1382–1402, Jul. 2011, doi: 10.1002/asi.21525.

[55] P. Pons and M. Latapy, "Computing communities in large networks using random walks," *J. Graph Algorithms Appl.*, vol. 10, no. 2, pp. 191–218, 2006, doi: 10.7155/jgaa.00124.

[56] A. M. Carrington, D. G. Manuel, P. W. Fieguth, T. Ramsay, V. Osmani, B. Wernly, C. Bennett, S. Hawken, O. Magwood, Y. Sheikh, M. McInnes, and A. Holzinger, "Deep ROC analysis and AUC as balanced average accuracy, for improved classifier selection, audit and explanation," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 45, no. 1, pp. 329–341, Jan. 2023, doi: 10.1109/TPAMI.2022.3145392.

[57] S. Wang and X. Yao, "Using class imbalance learning for software defect prediction," *IEEE Trans. Rel.*, vol. 62, no. 2, pp. 434–443, Jun. 2013, doi: 10.1109/Tr.2013.2259203.

[58] Microsoft. (2023). *Telemetry for Predictive Maintenance*. Accessed: Mar. 23, 2023. [Online]. Available: https://www.kaggle.com/datasets/arnabbiswas1/microsoft-azure-predictive-maintenance

[59] Pump. (2023). *Pump Sensor Data*. Accessed: Mar. 23, 2023. [Online]. Available: https://www.kaggle.com/datasets/nphantawee/pump-sensor-data

[60] M. Das, R. Cui, D. R. Campbell, G. Agrawal, and R. Ramnath, "Towards methods for systematic research on big data," in *Proc. IEEE Int. Conf. Big Data*, Oct. 2015, pp. 2072–2081. [Online]. Available: https://ieeexplore.ieee.org/document/7363989/

[61] I. Martinez, E. Viles, and I. G. Olaizola, "Data science methodologies: Current challenges and future approaches," *Big Data Res.*, vol. 24, May 2021, Art. no. 100183, doi: 10.1016/j.bdr.2020.100183.

[62] F. Kamalov and H. Sulieman, "Time series signal recovery methods: Comparative study," in *Proc. Int. Symp. Netw., Comput. Commun. (ISNCC)*, Oct. 2021, pp. 1–5.

[63] L. Wijesekara and L. Liyanage, "Air quality data pre-processing: A novel algorithm to impute missing values in univariate time series," in *Proc. IEEE 33rd Int. Conf. Tools Artif. Intell. (ICTAI)*, Nov. 2021, pp. 996–1001.

[64] O. Caelen, "A Bayesian interpretation of the confusion matrix," *Ann. Math. Artif. Intell.*, vol. 81, nos. 3–4, pp. 429–450, Dec. 2017, doi: 10.1007/s10472-017-9564-8.

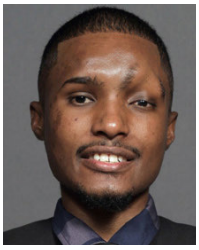
[65] S. Minaee, R. Kafieh, M. Sonka, S. Yazdani, and G. Jamalipour Soufi, "Deep-COVID: Predicting COVID-19 from chest X-ray images using deep transfer learning," *Med. Image Anal.*, vol. 65, Oct. 2020, Art. no. 101794, doi: 10.1016/j.media.2020.101794.

[66] D. K. McClish, "Analyzing a portion of the ROC curve," *Med. Decis. Making*, vol. 9, no. 3, pp. 190–195, Aug. 1989, doi: 10.1177/0272989x8900900307.

[67] D. J. Hand and R. J. Till, "A simple generalisation of the area under the ROC curve for multiple class classification problems," *Machine Learn.*, vol. 45, no. 2, pp. 171–186, Nov. 2001, doi: 10.1023/A:1010920819831.

[68] L. S. Shapley, *A Value for N-Person Games*. Santa Monica, CA, USA: RAND Corporation, 1952.

[69] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. 31st Int. Conf. Neural Inf. Process. Syst.*, Long Beach, CA, USA, 2017, pp. 1–10.
[70] C. Molnar, *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*, 1st ed. Victoria, BC, Canada: Leanpub, 2021.



MELVIN KISTEN received the B.Sc. degree (Hons.) in computer science from Sol Plaatje University (SPU), South Africa. He is currently pursuing the M.Sc. degree in computer science with North-West University (NWU), South Africa. He is also an M.Sc. Researcher with the Council for Scientific and Industrial Research (CSIR), South Africa. His primary research interests include software engineering and data science, with a particular emphasis on explainable artificial intelligence. His fervent engagement in various projects mirrors his commitment to these domains. This passion fuels his quest for deeper understanding, marking him as a dedicated scholar.

ABSALOM EL-SHAMIR EZUGWU received the B.Sc. degree in mathematics with computer science and the M.Sc. and Ph.D. degrees in computer science from Ahmadu Bello University, Zaria, Nigeria.

He is currently a Full Professor in computer science with the Unit for Data Science and Computing, North-West University, Potchefstroom, South Africa. He has contributed significantly to the academic community through the publication of numerous articles in internationally refereed journals, edited books, conference proceedings, and local journals. His research interests include artificial intelligence, swarm intelligence, and nature-inspired algorithm design, with a specific emphasis on computational intelligence and metaheuristic solutions for real-world global optimization problems. He is an active member of prominent organizations, such as Association for Computing Machinery (ACM), International Association of Engineers (IAENG), and Operations Research Society of South Africa (ORSSA). His dedication to advancing the field of computer science is evident in both his academic achievements and his ongoing contributions to cutting-edge research.

MICHEAL O. OLUSANYA (Member, IEEE) received the Ph.D. degree from the University of KwaZulu-Natal (UKZN), South Africa, in 2015. He is currently a Senior Lecturer with the Department of Computer Science and Information Technology, Sol Plaatje University (SPU), South Africa. His research interests include revolves around the application of metaheuristics and artificial intelligence techniques to solve real-life optimization problems, computational intelligence, and data analytics.