International Journal of Mechanical System Dynamics

WILEY

IJMSD International Journal of Mechanical System Dynamics Open Access

RESEARCH ARTICLE OPEN ACCESS

Safeguarding Pipeline Integrity Through Stacked Ensemble Learning and Data Fusion

Hussein A. M. Hussein[1,2] | Sharafiz B. Abdul Rahim[1,3] | Faizal B. Mustapha[4] | Prajindra S. Krishnan[5]

[1]Mechanical and Manufacturing Engineering, Universiti Putra Malaysia, Serdang, Malaysia | [2]Maintenance Department, Waha Oil Company, Tripoli, Libya | [3]Warwick Manufacturing Group, University of Warwick, UK | [4]Aerospace Engineering, Universiti Putra Malaysia, Serdang, Malaysia | [5]Electrical and Electronics Engineering, Universiti Tenaga Nasional, Kajang, Malaysia

Correspondence: Sharafiz B. Abdul Rahim (sharafiz@upm.edu.my)

Received: 7 October 2024 | **Revised:** 27 November 2024 | **Accepted:** 29 November 2024

Keywords: artificial neural networks (ANNs) | data fusion | ensemble learning | pipeline integrity | stacking method

ABSTRACT

This research presents a novel approach to pipeline Structure Health Monitoring (SHM) by utilizing frequency response function signals and integrating advanced data-driven techniques to detect and evaluate vibration responses regarding loose bolts, scale deposits within pipelines, and cracks at pipeline supports, aiming to determine the effectiveness of utilizing artificial neural networks (ANN) and an ensemble learning approach in detecting the aforementioned damages through a data-driven approach. The research starts by recording 6500 samples captured by two accelerometers, related to 11 replicated pipeline structural scenarios. The research demonstrated the potential of principal component analysis (PCA) in dimensionality reduction, achieving approximately 81% reduction in data set 1 acquired by accelerometer 1 and around 79.5% in data set 2 acquired by accelerometer 2, without significant loss of information. Additionally, two ANN base models were employed for fault recognition and classification, achieving over 99.88% accuracy and mean squared error values ranging from 0.00006 to 0.00019. A significant innovation of this work lies in the implementation of an ensemble learning approach, which integrates the strengths of the base models, showcasing outstanding performance that was proved consistent across multiple iterations, effectively mitigating the weaknesses of the base models and providing a reliable fault classification and prediction system. This research underscores the effectiveness of combining PCA, ANN, k-fold cross-validation, and ensemble learning techniques in pipeline SHM for improved reliability and safety. The findings highlight the potential for broader applications of this methodology in real-world scenarios, addressing urgent challenges faced by infrastructure owners and operators.

1 | Introduction

Pipelines play a critical role in the transportation of fluids across vast distances, serving industries such as oil and gas, water supply, and chemical processing. Generally, structural damage identification is a vital area within structural health monitoring (SHM), playing a key role in ensuring the safe operation of structures. In recent decades, advancements in industrial manufacturing technology and precision instrumentation have led to significant progress in this field [1, 2]. Among the various SHM methods developed, vibration-based techniques have emerged as the most widely used [3]. The pipeline industry has experienced substantial advancements, driven by factors such as sustainability, automation, enhanced performance, reliability, and the implementation of SHM techniques, which have become increasingly prevalent across various engineering sectors [4]. A notable advancement today is artificial neural networks (ANNs), a branch of machine learning that has been utilized across various tasks in multiple sectors of the economy, society, and everyday life. Unlike traditional computer programs, ANNs are tools that are trained using input–output data [5, 6]. They have demonstrated their ability to tackle numerous prediction challenges, although their accuracy depends heavily on the quality of the training data [7]. These developments have transformed the way pipelines are designed, operated, and maintained, leading to more efficient, environmentally conscious, and resilient infrastructure. Ensuring the integrity of pipelines is essential to prevent leaks, ruptures, and other forms of failure that can lead to catastrophic consequences, in terms of environmental damage, financial losses, and possible loss of lives. Traditional pipeline monitoring methods, relying on individual sensors or isolated inspection techniques, often fall short of providing a comprehensive understanding of pipeline conditions. This limitation has led to increased interest in developing advanced monitoring strategies that offer more holistic and accurate assessments of pipeline integrity. Conventional SHM techniques typically use single-sensor approaches, which may not adequately capture the complex nature of pipeline degradation and failure mechanisms. While these traditional methods have made progress, there is a growing need for more sophisticated techniques to effectively manage the complex, nonlinear relationships in pipeline sensor data. Enhanced monitoring approaches would enable a more accurate evaluation of overall pipeline integrity and condition, ultimately improving safety and reliability.

1.1 | Ensemble Learning

Ensemble learning provides novel approaches to address the challenges associated with traditional SHM techniques. Ensemble learning enhances predictive performance by leveraging the collective power of multiple models, addressing the shortcomings of traditional techniques that depend on individual model performance. Xiong et al. [8] recently highlighted the limitations of traditional SHM techniques, which often rely on manual interpretation and can compromise monitoring efficiency. The concept of ensemble learning, which combines two or more machine learning models, holds immense promise for enhancing the comprehensive assessment of the structure's integrity. Ensemble learning approaches leverage the strengths of multiple individual models to create a more robust and accurate overall prediction or decision-making system. The ensemble methods have been extensively explored across various disciplines, leading to a rich variety of techniques available for researchers and practitioners, and the growing interest in ensemble learning underscores its importance in enhancing classification accuracy and robustness in machine learning applications [9]. Bagging, boosting, stacking, and majority voting are the most common methods of ensemble learning and are widely used in machine learning competitions and real-world applications due to their ability to enhance model robustness and accuracy. Aiming to provide further information about the advantages of using multiple classifiers to enhance outcomes in machine learning tasks, Shah et al. [10] and Rincy and Gupta [11] discussed ensemble learning techniques in machine learning, highlighting their effectiveness in improving predictive performance and extensively explaining the bagging, boosting, and stacking methods, and compared the factors involved in these techniques. Boosting methods utilize an ensemble of multiple decision trees instead of relying on a single one, leading to a collective prediction. This approach combines several weak learners to create a more robust learner. The trees are arranged in a sequential manner, with each tree focusing on reducing the errors made by the previous tree. Due to this interconnected sequence, boosting algorithms may take longer to train but deliver highly accurate results [12]. Stacking is an ensemble machine learning algorithm that combines predictions from multiple diverse models. Unlike bagging, where the models are typically homogeneous, stacking employs different base models (level-0 models) to capture various aspects of the problem. A meta-model (level-1 model) is then used to learn how to best combine these predictions, rather than sequentially correcting errors like in boosting [13]. The configuration of a stacked model could involve the integration of various classification techniques, such as decision trees, random forests, support vector machines, and neural networks. Each of these models may excel at different aspects of the problem, and by combining their outputs, the ensemble can provide a more holistic and reliable assessment of the structure's condition and potential degradation or failure mechanisms.

Innovative ensemble learning approaches can greatly enhance pipeline SHM, improving the detection and prediction of potential issues. This enables proactive maintenance and risk mitigation, ultimately boosting pipeline safety and reliability. Despite their proven effectiveness in other fields, ensemble methods remain underutilized in pipeline integrity assessment. Exploring their application in SHM offers a promising research avenue, as these techniques could improve defect detection, data fusion, and predictive modeling, leading to more effective monitoring strategies for pipeline safety. One notable example in this regard is the use of data fusion in non-destructive evaluation (NDE) for pipeline defect identification, particularly utilizing the Learn++ algorithm [14]. Another application of ensemble learning in pipeline SHM is a framework that was proposed as an alternative approach for more efficient diagnosis in pipe damage condition monitoring [15]. The proposed framework utilized a bagging scheme, which combines multiple simple tree decision algorithms, to improve the accuracy of damage detection and classification. Warad et al. [16] recently presented an optimization ensemble learning-based model tailored for a large data set focused on water pipe leakage forecasting. Their proposed approach developed three distinct models, incorporating bagging, boosting, and Bayesian-based optimizable ensemble methods. The objective was to evaluate and select the model that demonstrated the most satisfactory performance for accurately predicting water leakage.

The limited adoption of ensemble learning in pipeline SHM is somewhat surprising, especially considering the inherent complexity and nonlinear relationships often present in pipeline sensor data. Ensemble learning methods are particularly well-suited for capturing these complexities, as they combine multiple models to improve predictive accuracy and robustness. Given the critical need for accurate defect detection and assessment in pipeline systems, leveraging ensemble techniques could significantly enhance monitoring capabilities. This gap highlights an opportunity for further research and application of ensemble learning in pipeline SHM, potentially leading to more effective maintenance strategies and improved infrastructure safety.

1.2 | Sensor Fusion

Sensor fusion is of paramount importance in the context of advanced pipeline SHM systems. Sensor fusion refers to the process of combining information obtained from various sensors, both of the same type and of different types, to create a more comprehensive and accurate understanding of the monitored system. In the case of pipeline integrity, sensor fusion can involve the integration of data from sensors positioned at different locations along the pipeline, as well as the integration of diverse sensor modalities. Pipeline systems are inherently complex and span vast areas, making comprehensive monitoring challenging. To address localized vulnerabilities, these large systems can be partitioned into smaller, manageable subsystems, similar to the approach used in finite element analysis for complex engineering problems. This targeted monitoring can be enhanced by the integration of renewable energy solutions and advanced data transmission technologies, allowing even remote sections of the pipeline systems to be effectively monitored. In this context, data fusion becomes crucial, as it integrates diverse data sources to provide a nuanced understanding of the pipeline's condition, accounting for the wide range of potential degradation and failure mechanisms that may not be fully captured by a single sensor or a limited set of sensors. Highlighting the importance of integrating multiple sensors to improve the accuracy of monitoring, Sharma et al. [17]. discussed the potential for using multisensor systems and sensor data fusion for condition-based maintenance with the great contribution of the advancements in electronics, sensor technology, information science, and electrical and computer engineering to the emerging technologies in pipeline SHM. Integrating sensor data, from the same type of sensors at different locations or from various sensor modalities, can uncover patterns, correlations, and interdependencies that individual readings might miss. This comprehensive view facilitates the detection of subtle changes, identification of emerging issues, and prediction of potential failures, ultimately enhancing pipeline integrity and safety. Implementing sensor fusion techniques in pipeline monitoring and SHM systems significantly contributes to advanced safeguarding strategies, providing a more holistic assessment of pipeline condition and performance. This approach supports the industry's focus on sustainability, automation, and reliability, enabling asset owners and operators to make informed decisions and adopt proactive maintenance measures. In summary, while ensemble learning can be applied to a single sensor data set, its effectiveness greatly improves with diverse sensor data integration through sensor fusion techniques, leading to more accurate and insightful predictions in pipeline SHM.

1.3 | Motivation and Expected Impact

This article introduces a Meta-ANN model combined with sensor fusion techniques to improve the accuracy and reliability of pipeline integrity assessments. It addresses gaps in existing methods and offers an innovative approach to safeguarding critical infrastructure. Conducted at the Sound and Vibration Research Group (SVRG) at University Putra Malaysia (UPM), the research leverages ensemble ANN (stacked ANN) models to enhance pipeline SHM using data-driven approaches. The study responds to the increasing frequency of pipeline failures due to various faults by utilizing a test setup that simulates three key fault modes: looseness of bolts, scale deposits, and cracks at supports. Monitoring these faults is crucial, as undetected issues can lead to severe consequences. The controlled experimental platform will generate data to train and validate the stacked ANN model, demonstrating its effectiveness compared to single-sensor ANN models. Successful implementation could significantly impact pipeline SHM, leading to advanced monitoring systems, improved asset management, reduced maintenance costs, and enhanced safety for operators and communities.

In conclusion, the article is structured as follows: Section 1 covers the background and motivation, Section 2 outlines the methodology, Section 3 discusses results and analysis, and Section 4 concludes with key insights, limitations, and recommendations, all supported by relevant images and tables for clarity.

2 | Research Methodology

This current work is built upon the authors' recently published work [18], where the experiment setup was thoroughly explained. In brief, the experimental work commenced with the setup of a controlled experimental environment as illustrated in Figure 1. This setup involved a pipeline segment on which three specific fault conditions, as depicted in Table 1, have been replicated. To acquire the frequency response functions (FRFs) on the pipe segment to investigate the effects of the replicated damages, piezoelectric accelerometers, described in Table 2, were positioned and affixed on the pipe segment and wire-connected to the Siemens LMS SCADAS Mobile data acquisition unit. The pipeline segment was excited within the flow-induced-vibration frequency range by a 24-DC-volt 4000-rpm electric vibrator with a speed controller and adjustable exciting force. In pursuit of the research objectives during the spectral testing, 1500 samples (runs) captured by two sensors, Input 2 and Input 3, were initially recorded for the fault-free scenario. Subsequently, 2000 samples were recorded for scenarios involving four levels of bolt looseness, with 500 samples for each severity. Following this, 1500 samples were recorded for three scale-formation scenarios, again with 500 samples for each severity. Finally, 1500 samples were recorded for three scenarios involving cracked support, with 500 samples for each severity. The outcome was two complete sensor data sets transferred from the LMS SCADAS Mobile unit to a PC, organized in two individual matrices each of 6500 rows (representing the number of samples) by 4096 columns (representing the number of features) related to 11 pipeline structural scenarios. While the data sets are limited to specific failure types, some degree of generalization can be theoretically considered. The underlying principles of how different failures affect pipeline integrity, such as changes in stiffness or mass, suggest that the model may still recognize patterns indicative of similar issues such as cracks, corrosion, and erosion. These data sets were saved and analyzed using a 12th Gen Intel® Core™ i5-1235U processor at 1.30 GHz, with 8.00 GB of RAM, a 64-bit operating system, and Windows 11, in the MATLAB® software environment version R2021b. The 64-bit operating system supports this by allowing MATLAB to utilize more memory, facilitating the execution of complex algorithms. By

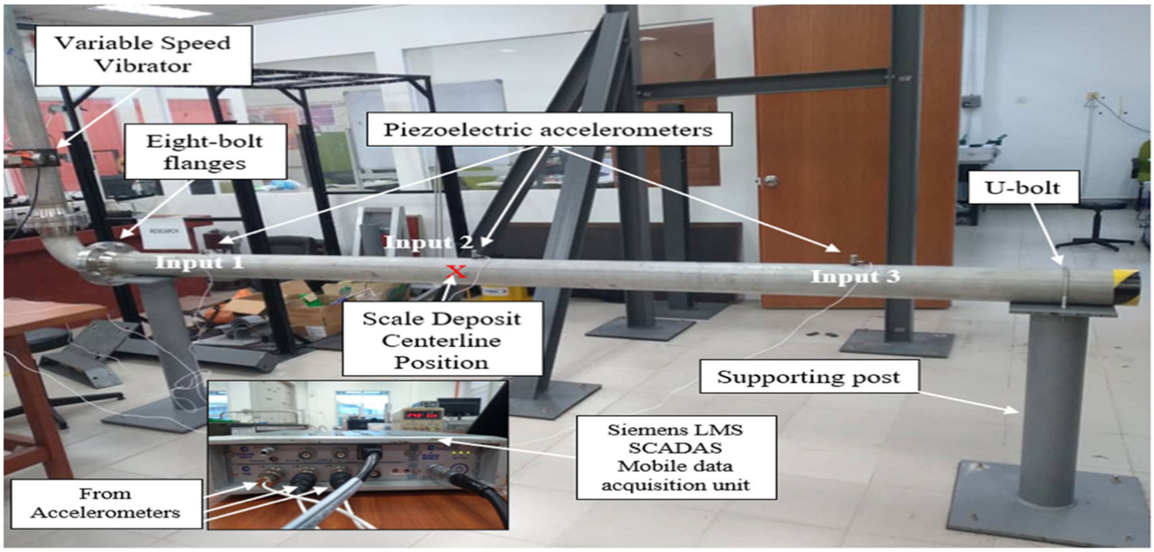

FIGURE 1 | Schematic of the experimental setup.

TABLE 1 | Description and designation of the replicated pipeline failures.

Failure mode	Severity level	Failure description	Failure designation
Bolt-loosening at connecting flanges	1st	Loosening of Bolt 1	BL-S1
	2nd	Loosening of Bolts 1 and 2	BL-S2
	3rd	Loosening of Bolts 1, 2, and 3	BL-S3
	4th	Loosening of Bolts 1, 2, 3, and 4	BL-S4
Scale deposit formation within pipelines	1st	13-gram PVC pipe segment	SD-S1
	2nd	130-gram PVC pipe segment	SD-S2
	3rd	520-gram PVC pipe segment	SD-S3
Crack occurrence at pipeline supports	1st	¼ Sawcut in U-bolt Support	CO-S1
	2nd	½ Sawcut in U-bolt Support	CO-S2
	3rd	Full Sawcut in U-bolt Support	CO-S3

TABLE 2 | Description and specifications of the used piezoelectric accelerometers.

Manufacturer	Designation	Measured quantity	Input mode/unit	Measured parameter	Actual sensitivity
PCB Piezotronics	Input 1 (Excitation)	Vibration	AC voltage (mV)	Acceleration	10.32 mV(m/s^2)
HBK	Input 2 (Response)	Vibration	AC voltage (mV)	Acceleration	1032.4 mV/g
HBK	Input 3 (Response)	Vibration	AC voltage (mV)	Acceleration	1008.2 mV/g

using MATLAB R2021b, built-in functions and the Parallel Computing Toolbox to utilize GPU capabilities can be leveraged, significantly accelerating processing times for data-intensive tasks.

This research implemented several key preprocessing techniques to enhance the quality of raw sensor data. Bandwidth settings were optimized for data acquisition, using a 2048 Hz bandwidth with 4096 spectral lines, resulting in a frequency resolution of 0.5 Hz. This ensured that the data captured the relevant frequencies for analysis. A uniform windowing function was applied to the captured signals to improve spectral data quality and extract meaningful frequency information from both reference and response signals. This technique mitigates spectral leakage and enhances the accuracy of subsequent analyses. Additionally, the initial and final segments of the acquired signals were removed to eliminate transient effects and noise distortions. This process produced two sub-matrices of dimensions [6500 × 3000], ready for principal component analysis (PCA), as shown in Figure 2.

The PCA [19] was applied to the produced 3000-dimensional data sets and sought to identify the number of principal components that captured at least 95% of the cumulative explained variance. The

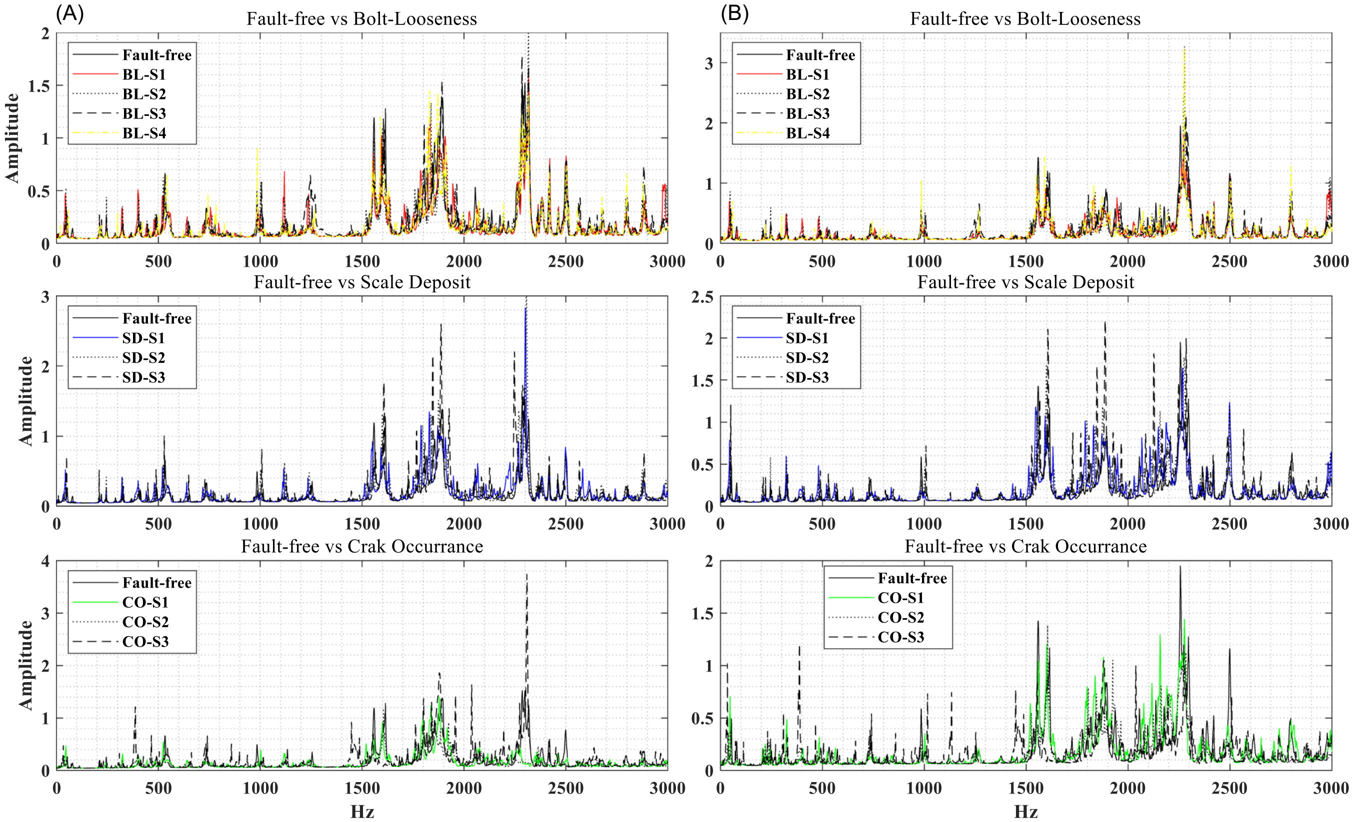

FIGURE 2 | Average frequency response function signals for 11 case studies acquired from two accelerometers. (A) From Input 2 and (B) from Input 3.

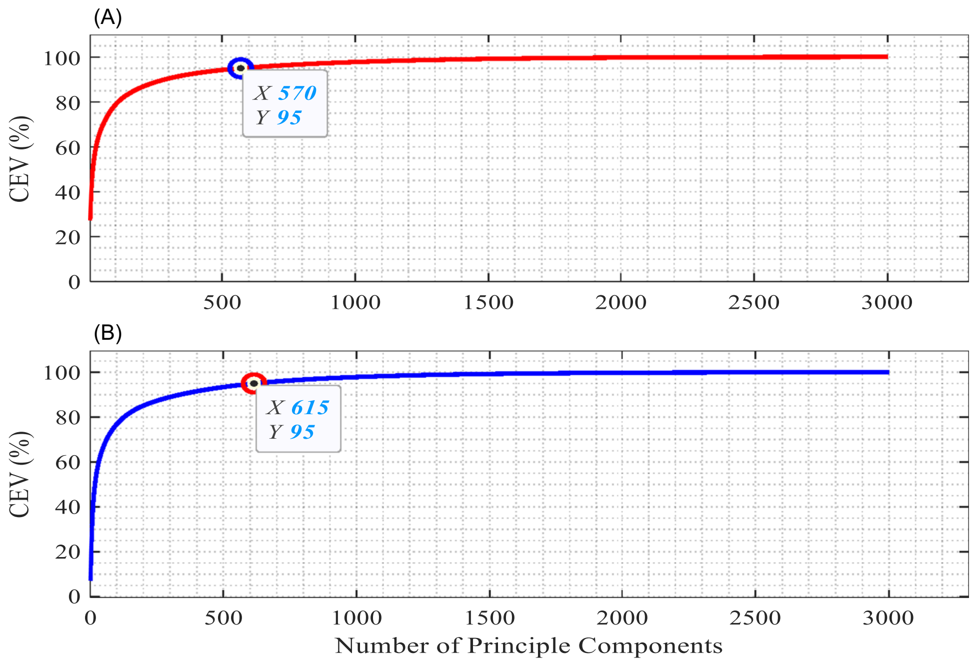

FIGURE 3 | Cumulative explained variance (CEV) by principal components. (A) For data set 1 and (B) for data set 2.

application of PCA to the produced data sets revealed key insights into their intrinsic dimensionality. As depicted in Figure 3, PCA analysis indicated that 570 principal components for Data set 1 and 615 for Data set 2 captured 95% of the cumulative explained variance. This means the data sets can be reduced to 570 and 615 features, respectively, achieving approximately 81% and 79.5% dimensionality reduction while retaining 95% of the variance. Such achievement is a substantial accomplishment and this level of reduction not only simplifies the data structure but also enhances computational efficiency, which is crucial for subsequent analyses and modeling tasks. Additionally, the retained 95% of cumulative explained variance indicates that the most significant patterns and structures in the data are preserved. While it is acknowledged that some subtle features may be lost, the primary focus of PCA is to maintain the overall data integrity and variance, which are essential for effective fault detection. However, the appropriate cumulative explained variance percentage is carefully selected to avoid missing informative data. The preserved variance reflects the dominant characteristics that influence the outcomes aimed to be analyzed.

Following the application of PCA for dimensionality reduction, the authors utilized MATLAB's *patternnet* function to develop ANN models. Before the application of the developed ANN models on the pre-processed data sets, the model's hyperparameters were carefully selected. Effective hyperparameter selection is crucial in optimizing machine learning models, as it directly influences their performance and predictive accuracy. Choosing the right hyperparameter values can significantly reduce prediction errors and enhance the model's ability to generalize to unseen data [20, 21]. In the authors' recent work [18, 22] an explanation of hyperparameter selection strategy was provided, detailing the methodology employed and its impact on the performance of the developed ANN model.

In this research, an ensemble learning model utilizing stacked generalization was employed, where two base models (ANN models) were trained to capture different patterns in the data. The inputs to these base models were the principal components selected from data sets 1 and 2, and the outputs were the predictions obtained from their respective data sets of the 11 replicated scenarios. A third ANN model that maintains the same hyperparameters as the base models was utilized as the meta-model at the top layer for final classification. The integration, illustrated in Figure 4, involved feature stacking utilizing decision-level fusion, where the outputs from each base-

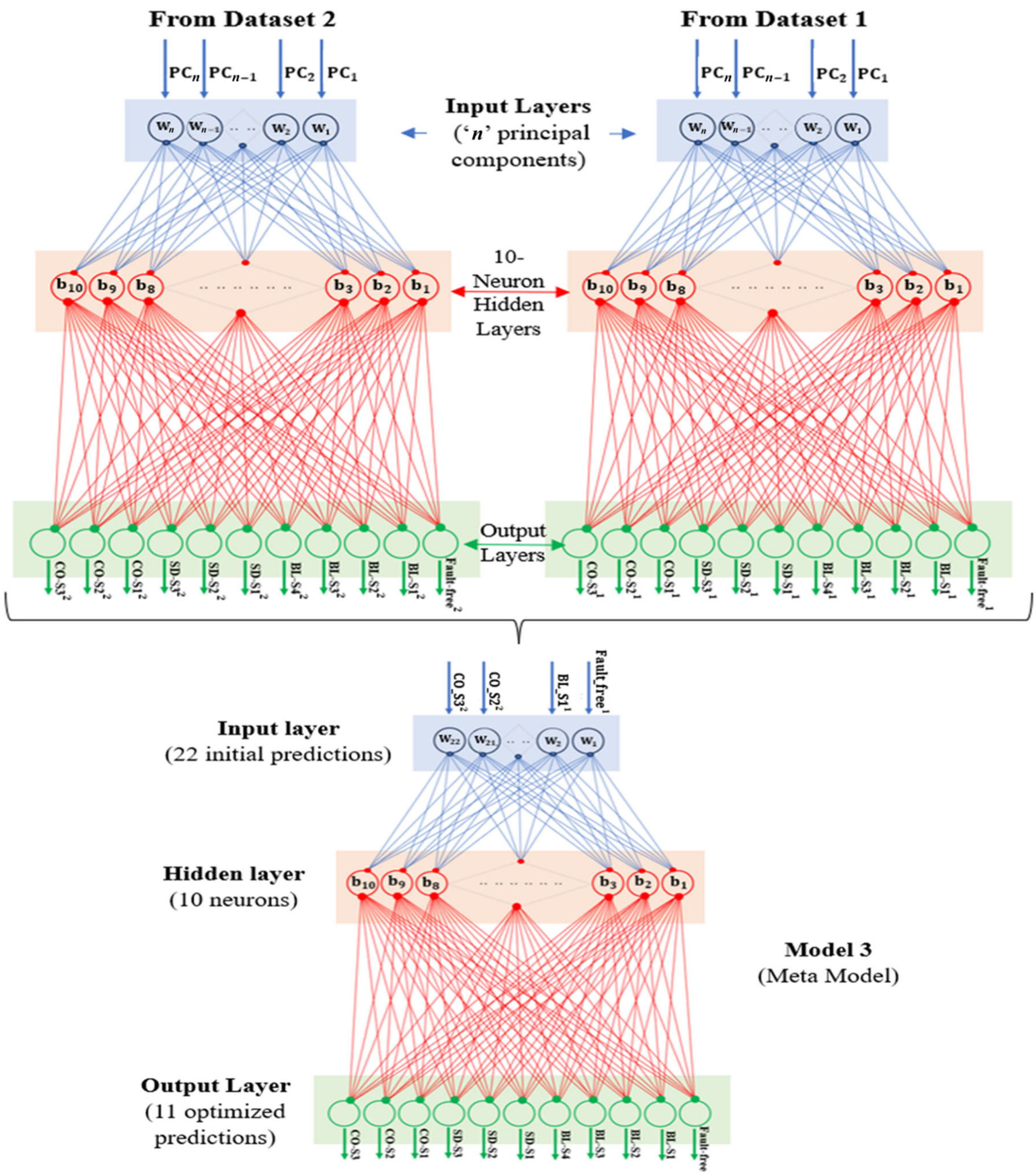

FIGURE 4 | Configuration of the ANN ensemble learning technique.

model served as input features for the meta-model. This meta-model was trained on these predictions, enabling it to learn how to best combine them to improve overall accuracy and robustness. The use of a third ANN allows for adaptive learning from the outputs of the base models, optimizing the combination of predictions based on the specific characteristics of the data. By employing an ANN as a meta-model, its ability to minimize errors through backpropagation is leveraged, potentially leading to better performance. Furthermore, using an ANN throughout (both as base- and meta-model) creates a cohesive framework that simplifies the training and, if required, the hyper-tuning process.

To evaluate the generalization of the proposed model to an independent data set, *K*-fold cross-validation (CV) was employed, which is crucial for assessing predictive power. This method divides the data set into *K* subsets or "folds," typically set to 10, although it can be adjusted based on the data set size [23, 24]. The model is trained *K* times, using $K-1$ folds for training and the remaining fold for testing, ensuring each data point is included in a test set once. Performance metrics (such as accuracy and mean squared error [MSE]) are calculated for each fold, with the final performance reported as the average across all *K* folds. *K*-fold CV provides a reliable estimate of model performance, maximizing both training and testing data while helping to understand bias and variance. For the proposed monitoring method, a 10-fold CV was implemented using MATLAB's *cvpartition* function to create the folds and *cv.training* and *cv.test* functions to generate training and testing indices before applying the *patternnet* function for pattern recognition of the replicated pipeline failures.

3 | Results and Discussion

With the completion of data collection, preprocessing, and analysis, the authors presented key findings from their investigation into pipeline infrastructure health monitoring. Using FRF signals from accelerometers and the pattern recognition capabilities of ANNs, the research addressed the challenge of enhancing pipeline monitoring for three common fault scenarios: loose bolts at flanges, scale deposits, and cracks at supports. The results highlighted the transformative role of data-driven techniques in pipeline SHM, demonstrating the effectiveness of their approach in accurately detecting and predicting potential failures. This discussion provided valuable insights into the practical implementation of the data-driven SHM methodology, addressing urgent challenges faced by infrastructure owners in preserving pipeline integrity. To ensure the robustness of the ANN ensemble model, the authors recognized that results from a single run could be misleading due to factors like random weight initialization. Therefore, they ran the model 10 times, implementing a 10-fold CV in each run to assess performance consistency. This repeated testing confirmed that the reported performance metrics accurately reflected the model's true capabilities. The performance evaluation of the ANN sub-models and the meta-model across these iterations revealed important insights into the effectiveness of the proposed approach in pipeline SHM.

The results in Table 3 indicate that while models trained with 10-fold CV perform well, the meta-model's performance appears less pronounced compared to the results in Table 4. This surprising finding can be attributed to several factors. The 10-fold CV is renowned for its robustness in assessing machine learning model performance by mitigating overfitting and providing a more generalized evaluation. However, it can also introduce variability due to the partitioning of the data set, which may affect performance metrics. In contrast, training without a 10-fold CV allows the model to learn from the entire data set, potentially leading to better performance metrics if the model inadvertently learns from the test data. While 10-fold CV ensures evaluation of unseen data, revealing true generalization capabilities, training without it risks overfitting, where the model captures noise or specific patterns that do not generalize well. Additionally, the absence of CV may yield seemingly better performance metrics that reflect the model's fit to the entire data set, which can be misleading regarding its performance on new data. Moreover, without a 10-fold CV, the model may benefit from favorable random weight initialization, while each fold in the CV may have different initializations,

TABLE 3 | Performance evaluation of the base models and meta-model across 10 iterations with 10-fold cross-validation.

	Base-model 1		Base-model 2		Meta-model	
Iteration number	**MSE**	**Overall accuracy**	**MSE**	**Overall accuracy**	**MSE**	**Overall accuracy**
1	0.00020	99.862	0.00052	99.631	0.00010	99.938
2	0.00030	99.800	0.00042	99.677	0.00012	99.923
3	0.00036	99.738	0.00040	99.692	0.00013	99.892
4	0.00016	99.877	0.00070	99.538	0.00007	99.954
5	0.00027	99.831	0.00055	99.600	0.00017	99.908
6	0.00019	99.892	0.00198	98.538	0.00015	99.908
7	0.00020	99.862	0.00066	99.523	0.00006	99.954
8	0.00024	99.846	0.00062	99.554	0.00019	99.877
9	0.00031	99.785	0.00045	99.708	0.00018	99.908
10	0.00045	99.692	0.00059	99.600	0.00013	99.923

Abbreviation: MSE, mean squared error.

TABLE 4 | Performance evaluation of the base models and meta-model across 10 iterations without a 10-fold cross-validation.

	Base-model 1		Base-model 2		Meta-model	
Iteration number	MSE	Overall accuracy	MSE	Overall accuracy	MSE	Overall accuracy
1	0.00004	99.985	0.00009	99.936	0	100
2	0.00004	99.954	0.00008	99.936	0	100
3	0.00002	99.985	0.00009	99.923	0	100
4	0.0001	99.934	0.00014	99.892	0.00006	99.954
5	0	100	0.00007	99.939	0.000004	100
6	0.00008	99.9385	0.00016	99.877	0.00006	99.954
7	0.00010	99.9231	0.0001	99.939	0.00006	99.954
8	0.00005	99.9846	0.00017	99.877	0	100
9	0.00008	99.9538	0.00008	99.954	0	100
10	0.00001	99.9846	0.00021	99.862	0.00006	99.954

Abbreviation: MSE, mean squared error.

leading to performance variations. The significant performance difference observed without a 10-fold CV suggests that the meta-model appears superior; however, this may not hold true in practical applications. These findings underscore the importance of employing robust evaluation techniques like CV to ensure that reported performance metrics accurately reflect true model capabilities. Ultimately, this analysis highlights a critical trade-off between generalization, achieved through a 10-fold CV, and the potential for improved predictive accuracy when training on the entire data set, necessitating careful consideration in model training strategies. In summary, the obtained results indicate that while the meta-model shows promise when evaluated without the CV, its performance does not significantly outperform the base models when assessed more rigorously with a 10-fold CV. This emphasizes the need for careful evaluation of model performance to avoid overestimating effectiveness based on potentially misleading metrics. It is crucial to rely on CV results for a more accurate assessment of model generalization.

A detailed breakdown of the results in Table 3, reflecting the performance of both base models and the meta-model under 10-fold CV, provides valuable insights into the effectiveness of the proposed monitoring technique. The assessment of metrics such as MSE and overall accuracy reveals the strengths and weaknesses of the models. Training the stacked ensemble monitoring system across 10 iterations with a 10-fold CV enables a comprehensive evaluation of robustness and consistency, critical for pipeline integrity monitoring. Base-model 1 exhibits high robustness, with MSE values ranging from 0.000 16 to 0.000 45 and accuracy consistently above 99.7%. In contrast, base-model 2 shows slight variability, particularly in iteration 6, where accuracy dipped to 98.538%, though it maintains generally low MSE values. The meta-model demonstrates superior performance, achieving MSE as low as 0.000 06 and accuracy reaching 99.954% in iteration 4. This indicates its effectiveness in aggregating predictions from the base models, enhancing overall prediction quality. Notably, the meta-model consistently achieves lower MSE values than both base models; for instance, in iteration 4, its MSE is 0.000 07 compared to 0.000 16 for base-model 1 and 0.000 70 for base-model 2. Additionally, the meta-model often surpasses the accuracy of the base models, reaching 99.938% in iteration 1 versus 99.862% for base-model 1 and 99.631% for base-model 2. Overall, the evaluation highlights the robustness of the proposed models, particularly the meta-model, which consistently outperforms both base models. This underscores the effectiveness of ensemble methods in enhancing predictive performance and reliability, making a compelling case for their application in monitoring pipeline systems, where precision is crucial. It is acknowledgeable that achieving such strong performance metrics in controlled experimental conditions can be questionable and raise some concerns. However, these controlled experiments serve to establish a baseline for the proposed methods' effectiveness under specific fault scenarios. While the results may seem overly optimistic, they highlight the model's ability to learn distinct patterns. This pilot study is essential, as it isolates variables and assesses performance without real-world confounding factors, providing valuable insights before applying the proposed method in more complex, real-world settings.

To demonstrate statistically significant differences in performance metrics (MSE and overall accuracy) between the base models and the meta-model, a t-test was employed. This test assesses the means of two groups, revealing whether significant differences exist. The hypothesis variable (h) indicates rejection of the null hypothesis ($h = 1$) or failure to reject it ($h = 0$). A p-value $< .05$ signifies a statistically significant difference. The test results indicate significant differences in both MSE and accuracy among the models, underscoring their varying effectiveness. Figures 5 and 6 display heatmaps of the p-values from the t-tests, organized in a 3×3 grid. Diagonal cells are marked as "NaN," reflecting self-comparisons. The color gradient illustrates significance: darker colors indicate higher p-values (no significant difference), while brighter colors highlight lower p-values (significant differences). The heatmaps effectively visualize these statistical differences, aiding in the interpretation of the t-test results. The t-test results indicate significant prediction improvements provided by the meta-model compared to the base models across both metrics. These results confirm that the meta-model consistently outperforms both base-model 1 and base-model 2 in terms of both MSE and overall

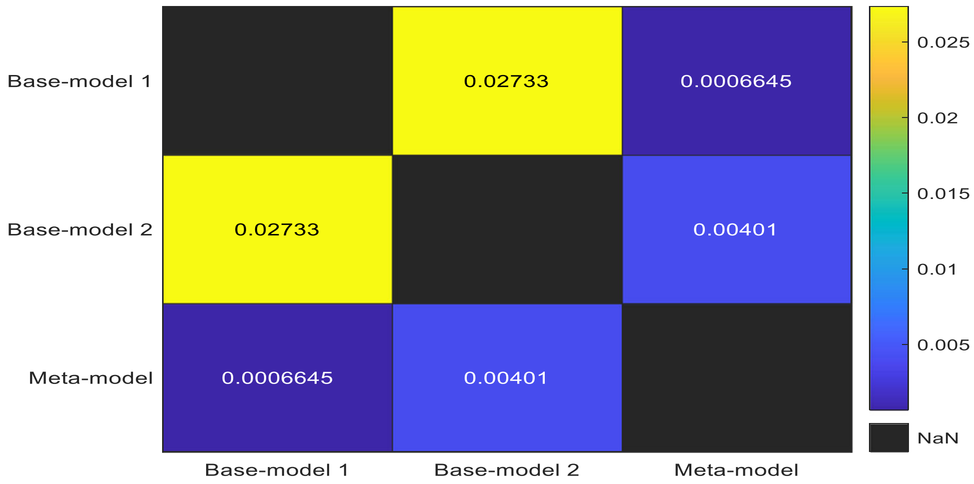

FIGURE 5 | Heatmap of *p*-values from *t*-test on mean squared error.

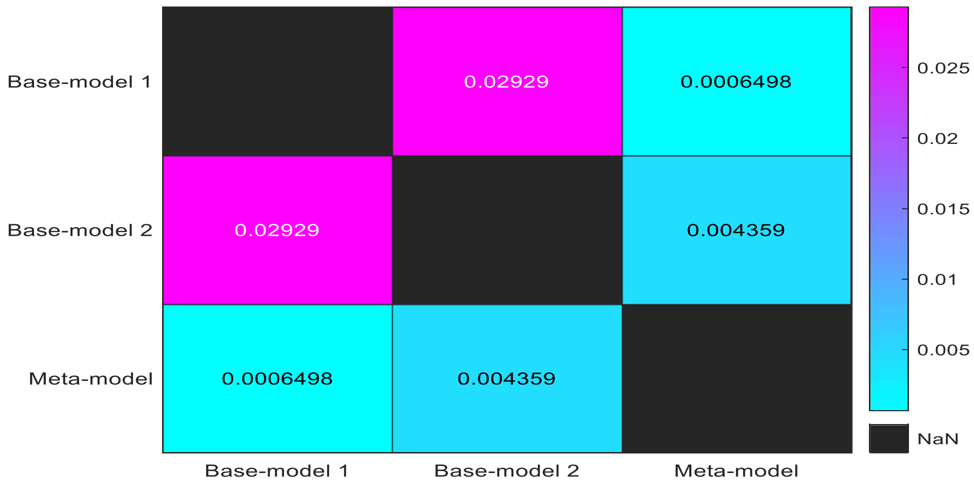

FIGURE 6 | Heatmap of *p*-values from *t*-test on overall accuracy.

TABLE 5 | Confusion matrix for the 10th iteration of the sub-model 1.

150	0	0	0	0	0	0	0	0	0	0
0	50	0	0	0	0	0	0	0	0	0
0	0	47	3	0	0	0	0	0	0	0
0	0	1	49	0	0	0	0	0	0	0
0	0	0	0	50	0	0	0	0	0	0
0	0	0	0	0	50	0	0	0	0	0
0	0	0	0	0	0	50	0	0	0	0
0	0	0	0	0	0	0	50	0	0	0
0	0	0	0	0	0	0	0	50	0	0
0	0	0	0	0	0	0	0	0	50	0
0	0	0	0	0	0	0	0	0	0	50

accuracy, highlighting its effectiveness in enhancing predictive performance.

Furthermore, the classification performance of the models, as illustrated by the confusion matrices depicted in Tables 5–7, provides valuable insights into their strengths and weaknesses, particularly in relation to the specific fault scenarios they are designed to detect. Base-model 1 exhibits commendable accuracy, particularly in Class 1, which represents a fault-free scenario, where it achieved 150 true positives. This indicates that the model is highly effective at identifying instances without faults, which is critical for operational reliability. However, it

TABLE 6 | Confusion matrix for the 10th iteration of the sub-model 2.

150	0	0	0	0	0	0	0	0	0	0
0	50	0	0	0	0	0	0	0	0	0
1	0	48	1	0	0	0	0	0	0	0
0	0	1	49	0	0	0	0	0	0	0
0	0	0	0	50	0	0	0	0	0	0
0	0	0	0	0	50	0	0	0	0	0
0	0	0	0	0	0	50	0	0	0	0
0	0	0	0	0	0	0	50	0	0	0
0	0	0	0	0	1	0	0	49	0	0
0	0	0	0	0	0	0	0	0	50	0
0	0	0	0	0	0	0	0	0	0	50

TABLE 7 | Confusion matrix for the 10th iteration of the meta-model.

150	0	0	0	0	0	0	0	0	0	0
0	50	0	0	0	0	0	0	0	0	0
0	0	49	1	0	0	0	0	0	0	0
0	0	1	49	0	0	0	0	0	0	0
0	0	0	0	50	0	0	0	0	0	0
0	0	0	0	0	50	0	0	0	0	0
0	0	0	0	0	0	50	0	0	0	0
0	0	0	0	0	0	0	50	0	0	0
0	0	0	0	0	0	0	0	50	0	0
0	0	0	0	0	0	0	0	0	50	0
0	0	0	0	0	0	0	0	0	0	50

struggles with classes 3 and 4, both of which correspond to bolt looseness in connecting flanges. The misclassification of three instances of Class 3 as Class 4 suggests a potential overlap in the features that characterize these conditions, indicating that the model may not be adequately distinguishing between the severity of bolt looseness. This highlights a need for enhanced feature engineering or the inclusion of additional data to improve differentiation between these closely related fault types. In contrast, base-model 2 similarly demonstrates strong performance in Class 1 with perfect classification. However, it shows slightly more misclassification in the bolt-looseness categories, with one instance of Class 3 misclassified as Class 1 (a fault-free scenario) and another as Class 4. This indicates a higher false negative rate for Class 3 compared to base-model 1, suggesting that the model may be overly conservative in predicting faults in this category, potentially due to insufficient distinguishing features in training data. The meta-model stands out with exceptional performance. However, it exhibits two misclassifications: one in Class 3 and another in Class 4, both of which relate to bolt looseness. This indicates a false negative for each class, suggesting that while the meta-model improves upon the base models, there are still challenges in accurately identifying specific fault conditions related to bolt looseness. The misclassification of Class 3 and Class 4 implies that the model may struggle to differentiate between varying degrees or manifestations of bolt looseness, which could be due to overlapping features or insufficient training data specific to these classes. In summary, while all models exhibit strong classification capabilities, the analysis of true positives, false positives, and false negatives provides a nuanced understanding of their performance in the context of fault detection. Base-model 1 and base-model 2 excel in identifying fault-free scenarios but reveal areas needing improvement, particularly in distinguishing between the bolt-looseness classes (3 and 4). The meta-model's robust performance highlights its potential as a reliable tool for classification tasks, effectively minimizing misclassifications overall, yet it still faces challenges in accurately detecting specific fault conditions.

Following significant improvements in overall performance, MSE, and confusion matrices relative to traditional single-ANN models, the proposed stacked-ANN ensemble learning technique was thoroughly evaluated against established ensemble models in the pipeline SHM literature. This research aims to provide relevant insights for practitioners and researchers addressing similar challenges in pipeline defect detection. By focusing on the relatively established studies, the current work builds upon existing methodologies and performance benchmarks, offering a novel approach that enhances understanding in the field of pipeline SHM. This targeted analysis fills a

TABLE 8 | Comparison between the proposed ensemble method and other references.

Source of study	Methodology	Description	Performance
Ref. 14, 2004	Majority voting of Learn++ algorithm with the ensemble approach	The objective was to identify pitting, cracks, mechanical damage, and weld failures using real-world images captured by magnetic flux leakage (MFL) and ultrasonic testing (UT) techniques	Confidence interval of 95.022 ± 1.9985%
Ref. 15, 2017	Bagging of statistical modeling through PCA and piezo-diagnostics approach	The objective was to experimentally detect the position of an added mass on a pipeline's outer surface using piezoelectric devices	MSE < 0.8
Ref. 17, 2024	Optimization of bagging and boosting ensemble learning-based model	The objective was to forecast the water pipe leakage by tuning the hyperparameters of Bagging and boosting regression models using Bayesian optimization based on design, operation, and historical data.	MSE of 0.053E−04
Present work	Stacked ensemble learning. Two ANN-based models fed to an ANN-meta model	The objective is to experimentally detect bolt-looseness, scale deposit formation, and cracks in pipeline supports utilizing FRF	MSE < 1.90E−04 Overall accuracy > 99.88%

Abbreviations: ANN, artificial neural network; FRF, frequency response function; MSE, mean squared error; PCA, principal component analysis.

specific gap in the literature and advances the field effectively. The comparative analysis, illustrated in Table 8, contextualizes the proposed method's effectiveness by comparing its performance metrics with various benchmark models. This systematic assessment highlights the unique advantages and robustness of the stacked-ANN approach, underscoring its potential as a leading solution for pipeline defect detection.

Table 8 summarizes various studies employing ensemble learning methodologies to address specific engineering challenges in pipelines, highlighting the novelty and robustness of the present work compared to the referenced studies. While methods such as majority voting, bagging, and statistical modeling are established techniques, the stacking approach in the current work represents a more advanced ensemble strategy that can capture complex relationships in the data more effectively. The present work's emphasis on a diverse range of pipeline failures, including scale deposit formation, is novel and adds significant value to the field. Additionally, while other studies report competitive metrics (e.g., MSE < 0.8 for Ref. 15 and MSE of 0.053E−04 for Ref. 16), the present work's performance outshines these results, particularly in terms of overall accuracy. Furthermore, the confidence interval from Ref. 14 provides a measure of reliability, while the present work's direct accuracy measurement offers a clear picture of model effectiveness, which is crucial in real-world applications where data may vary significantly from training data sets.

The scalability of the proposed monitoring technique is noteworthy, as it partitions larger pipeline systems into smaller, manageable subsystems, aggregating their outcomes through ensemble learning. This modular approach enables effective monitoring across extensive networks, allowing adaptation to various pipeline sizes and complexities. The proposed data-driven method serves as a cornerstone for broader monitoring advancements in pipeline SHM. While the technique is FRF-based, it can dynamically adjust its protocols by sensor-fusing other modalities that measure temperature, pressure, and fluid properties, enhancing data robustness against dynamic environmental factors. To fully assess the adaptability of this stacking approach, scalability testing in larger and more complex pipeline systems is highly recommended. Additionally, while the present work was experimental in nature and due to the constrained resources and access to real-world pipeline facilities, it was challenging to conduct extensive field validation at this stage. However, the authors recognized the importance of validating their findings in practical environments to ensure robustness and applicability. To address this limitation, the full focus was on creating a comprehensive and controlled experimental framework that simulates critical fault conditions representative of realistic pipeline challenges. This pilot study allows for systematically exploring the effectiveness of the proposed method in a controlled setting before applying it to more complex scenarios.

4 | Conclusion

The importance of effective pipeline maintenance and monitoring is critical for ensuring safe and efficient operation of these vital infrastructure systems. Failures in pipeline systems can lead to severe safety hazards, environmental degradation, and significant economic losses. This study demonstrates that ensemble learning, particularly through the integration of multiple machine learning models, significantly enhances the assessment of pipeline integrity. By leveraging the strengths of various models, this approach provides more robust predictions. The use of PCA for dimensionality reduction, combined with ANNs, has proven effective in detecting and diagnosing pipeline faults, such as bolt looseness, scale deposits, and cracks. The proposed ensemble learning method shows promising results, achieving high accuracy and low MSE, indicating its practical applicability in real-world scenarios. The scalability of this method allows for effective monitoring across diverse pipeline systems, making it adaptable to varying sizes and complexities. Future work should focus on validating these findings in operational environments and conducting cost–benefit analyses to evaluate the economic impact of implementing this model. Overall, the integration of ensemble learning techniques represents a valuable advancement in pipeline maintenance and monitoring, contributing to enhanced infrastructure reliability and safety.

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

1. S. Rahim. "Investigating the Effect of Variable Mass Loading in Structural Health Monitoring From a Machine Learning Perspective," PhD thesis, University of Sheffield, 2018.

2. Y. Ma, F. Mustapha, M. Ishak, S. A. Rahim, and M. Mustapha, "Data-Driven Methods for Damage Detection and Identification of UAV: A Review," *Journal of Aeronautics, Astronautics and Aviation* 54, no. 4 (2022): 405–420.

3. Y. Li, H.-L. Minh, M. Cao, X. Qian, and M. Abdel Wahab, "An Integrated Surrogate Model-Driven and Improved Termite Life Cycle Optimizer for Damage Identification in Dams," *Mechanical Systems and Signal Processing* 208 (2024): 110986.

4. S. A. Rahim, G. Manson, and M. A. Aziz, "Data Clustering Based on Gaussian Mixture Model and Expectation-Maximization Algorithm for Datadriven Structural Health Monitoring System," *International Journal of Integrated Engineering* 13, no. 7 (2021): 167–175, https://publisher.uthm.edu.my/ojs/index.php/ijie/article/view/9540.

5. R. Zhao, S. H. Tang, E. E. Bin Supeni, S. A. Rahim, and L. Fan, "Z-YOLOv8s-Based Approach for Road Object Recognition in Complex Traffic Scenarios," *Alexandria Engineering Journal* 106 (2024): 298–311.

6. R. Zhao, S. Tang, E. E. B. Supeni, S. B. A. Rahim, and L. Fan, "A Review of Object Detection in Traffic Scenes Based on Deep Learning," *Applied Mathematics and Nonlinear Sciences* 9, no. 1 (2024).

7. L. Nguyen-Ngoc, Q. Nguyen-Huu, G. De Roeck, T. Bui-Tien, and M. Abdel Wahab, "Deep Neural Network and Evolved Optimization Algorithm for Damage Assessment in a Truss Bridge," *Mathematics* 12, no. 15 (2024): 2300.

8. C. Xiong, S. Lian, and W. Chen, "An Ensemble Method for Automatic Real-Time Detection, Evaluation and Position of Exposed Subsea Pipelines Based on 3D Real-Time Sonar System," *Journal of Civil Structural Health Monitoring* 13 (2023): 485–504.

9. L. Rokach, "Ensemble-Based Classifiers," *Artificial Intelligence Review* 33 (2010): 1–39.

10. M. Shah, K. Gandhi, K. A. Patel, H. Kantawala, et al., "Theoretical Evaluation of Ensemble Machine Learning Techniques," *5th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, 829-837, Tirunelveli, India, 2023.

11. T. N. Rincy and R. Gupta. "Ensemble Learning Techniques and its Efficiency in Machine Learning: A Survey". *2nd International Conference on Data, Engineering and Applications (IDEA)*, 1–6, Bhopal, India, 2020.

12. B. L. Dang, H. Nguyen-Xuan, and M. Abdel Wahab, "An Effective Approach for VARANS-VOF Modelling Interactions of Wave and Perforated Breakwater Using Gradient Boosting Decision Tree Algorithm," *Ocean Engineering* 268 (2023): 113398.

13. J. Brownlee, "Stacking ensemble machine learning With Python. Machine Learning Mastery, Ensemble Learning". Accessed September 3, 2024, https://machinelearningmastery.com/stacking-ensemble-machine-learning-with-python/.

14. D. Parikh, M. T. Kim, J. Oagaro, S. Mandayam, and R. Polikar, "Combining Classifiers for Multisensor Data Fusion," *IEEE International Conference on Systems, Man and Cybernetics (IEEE Cat. No.04CH37583)* (The Netherlands: The Hague, 2004). 2, 1232–1237.

15. J. Camacho-Navarro, M. Ruiz, R. Villamizar, L. Mujica, and G. Moreno-Beltrán, "Ensemble Learning As Approach for Pipeline Condition Assessment," *Journal of Physics: Conference Series* 842 (2017): 012019.

16. A. A. M. Warad, K. Wassif, and N. R. Darwish, "An Ensemble Learning Model for Forecasting Water-Pipe Leakage," *Scientific Reports* 14 (2024): 10683.

17. V. B. Sharma, K. Singh, R. Gupta, et al., "Review of Structural Health Monitoring Techniques in Pipeline and Wind Turbine Industries," *Applied System Innovation* 4, no. 3 (2021): 59.

18. H. A. M. Hussein, Sharafiz Rahim, F. B. Mustapha, P. S. Krishnan, and N. A. B. A. Jalil, "Data-Driven Multi-Fault Detection in Pipelines Utilizing Frequency Response Function and Artificial Neural Networks," *Journal of Pipeline Science and Engineering* (2024): 100223.

19. S. A. Rahim and G. Manson, "Kernel Principal Component Analysis for Structural Health Monitoring and Damage Detection of an Engineering Structure Under Operational Loading Variations," *Journal of Failure Analysis and Prevention* 21 (2021): 1981–1990.

20. N. Shawki, R. R. Nunez, I. Obeid, and J. Picone, "On Automating Hyperparameter Optimization for Deep Learning Applications," *IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, Philadelphia, PA, USA, 2021.

21. P. Nair, V. Vakharia, M. Shah, et al., "Ai-Driven Digital Twin Model for Reliable Lithium-Ion Battery Discharge Capacity Predictions," *International Journal of Intelligent Systems* 2024 (2024): 1–18.

22. Sharafiz Rahim, G. Manson, and F. Mustapha, "Damage Detection of the Jabiru's Aircraft Wing under Operational Fuel Loading Conditions Using Neural Network," *J Aeronaut Astronaut Aviat* 56, no. 1 (2024): 1–7.

23. V. Vakharia, V. K. Gupta, and P. K. Kankar, "A Comparison of Feature Ranking Techniques for Fault Diagnosis of Ball Bearing," *Soft Computing* 20 (2016): 1601–1619.

24. O. Oyedele, "Determining the Optimal Number of Folds to Use in a K-Fold Cross-Validation: A Neural Network Classification Experiment," *Research in Mathematics* 10, no. 1 (2023).