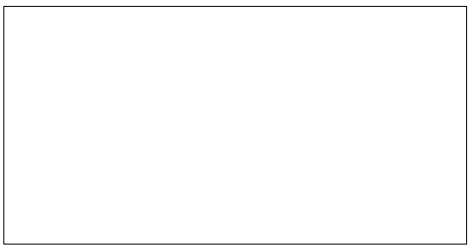


International Journal of Medical Informatics 206 (2026) 106162

Contents lists available at ScienceDirect

International Journal of Medical Informatics

journal homepage: www.elsevier.com/locate/ijmedinf

Hybridized artificial intelligence system for reducing neonatal mortality in Nigeria

Charity S. Odeyemi [a,b], Olatayo M. Olaniyan [b], Bolaji A. Omodunbi [b], Ibitoye B. Samuel [c], Afeez A. Soladoye [b,d], David B. Olawade [e,f,g,*]

[a] Department of Computer Engineering, School of Electrical Systems Engineering, Federal University of Technology, Akure, Nigeria
[b] Department of Computer Engineering, Federal University, Oye-Ekiti, Nigeria
[c] Department of Paediatrics, Federal Medical Centre, Owo, Nigeria
[d] Department of Computer Engineering, Adeleke University, Ede, Nigeria
[e] Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom
[f] Department of Research and Innovation, Medway NHS Foundation Trust, Gillingham ME7 5NY, United Kingdom
[g] Department of Public Health, York St John University, London, United Kingdom

ARTICLE INFO

Keywords:
Artificial intelligence
Hybrid model
Neonatal mortality
Deep learning
Healthcare diagnostics

ABSTRACT

Background: Neonatal diseases represent the leading cause of death in Nigeria, ranking the country second globally in neonatal mortality rates. Early and accurate diagnosis remains challenging, leading to delayed interventions and increased mortality.

Aim: To develop an artificial intelligence system capable of detecting multiple neonatal diseases using local datasets and advanced machine learning techniques to facilitate early intervention and reduce neonatal mortality in Southwest Nigeria.

Methods: Clinical records from 4,027 previously treated neonatal patients were collected from five tertiary hospitals across three Southwest Nigerian states. The dataset underwent comprehensive analysis, balancing using Synthetic Minority Over-sampling Technique (SMOTE), and preprocessing before training three deep learning models: Artificial Neural Networks (ANN), Long Short-Term Memory (LSTM), and a novel hybrid LSTM-ANN architecture. Model performance was evaluated using accuracy, precision, recall, and F1-score metrics with rigorous subject-wise validation and statistical testing.

Results: The hybrid LSTM-ANN model demonstrated superior performance with 82 % accuracy, 88 % precision, 82 % recall, and 86 % F1-score, significantly outperforming both standalone ANN (80 % accuracy) and LSTM (77 % accuracy). Disease-specific classification revealed exceptional performance for sepsis (precision: 0.90, F1-score: 0.88), birth asphyxia (0.88, 0.85), jaundice (0.86, 0.83), and prematurity (0.82, 0.80). McNemar's test confirmed significant hybrid superiority over ANN ($\chi 2 = 12.45$, $p < 0.001$) and LSTM ($\chi 2 = 15.67$, $p < 0.001$), whilst Friedman test ($\chi 2 = 18.42$, $p < 0.001$) validated the 5–6 % accuracy improvement.

Conclusion: The hybrid LSTM-ANN model establishes a valuable diagnostic tool for early neonatal disease detection. However, external validation and prospective clinical trials are necessary before clinical deployment.

1. Introduction

Machine intelligence technologies have fundamentally transformed how medical professionals approach diagnostic challenges across numerous clinical domains [1]. These computational systems replicate cognitive functions, enabling automated decision-making processes that achieve optimal outcomes through probabilistic reasoning [2] Within this technological landscape, machine learning algorithms demonstrate remarkable capacity to extract insights from complex datasets without explicit programming, whilst deep learning architectures employ multi-layered neural networks to accomplish sophisticated pattern recognition tasks [3]. Such systems process diverse information types, including visual, textual, and auditory data, to generate clinically relevant predictions through computational approaches that mirror biological

* Corresponding author at: Department of Allied and Public Health, School of Health, Sport and Bioscience, University of East London, London, United Kingdom.
E-mail addresses: charityodeyemi@gmail.com (C.S. Odeyemi), bolaji.omodunbi@fuoye.edu.ng (B.A. Omodunbi), gbemitoye91212@yahoo.com (I.B. Samuel), afeez.soladoye@fuoye.edu.ng (A.A. Soladoye), d.olawade@uel.ac.uk (D.B. Olawade).

https://doi.org/10.1016/j.ijmedinf.2025.106162
Received 2 October 2025; Received in revised form 16 October 2025; Accepted 17 October 2025
Available online 19 October 2025

neural processing [4].

Contemporary healthcare applications of intelligent algorithms have yielded significant diagnostic improvements, particularly for disease prediction in resource-constrained environments [5]. Stroke prediction models employing gated recurrent architectures have demonstrated considerable accuracy when properly validated using subject-specific methodologies in Sub-Saharan populations [6]. Similarly, coronary artery disease detection systems combining Random Forest classifiers with bio-inspired optimisation algorithms have substantially exceeded traditional risk assessment tools in predictive performance [7]. These developments underscore machine learning's potential to address critical healthcare delivery challenges where conventional diagnostic approaches prove insufficient. Hybrid methodologies integrating multiple algorithmic approaches have proven especially effective for complex medical classification tasks [8,9].

Nigeria confronts an urgent neonatal mortality crisis, with death rates during the first 28 days of life ranking among the world's highest [10,11]. This vulnerable developmental period exposes infants to numerous life-threatening conditions including sepsis, birth asphyxia, jaundice, and complications from premature delivery [12]. Diagnostic uncertainties and delayed clinical recognition contribute substantially to preventable deaths, making accurate early detection paramount for survival [13]. Whilst existing research has examined mortality determinants [14–16] and maternal health factors [17–19], integrated AI classification systems trained on local clinical data remain absent from Nigerian healthcare contexts. Current diagnostic protocols typically identify diseases only after significant pathological progression, when therapeutic interventions become less effective.

This research addresses a critical knowledge gap by developing Nigeria's first hybrid deep learning system for simultaneous multi-disease neonatal classification. Previous AI implementations in well-resourced settings demonstrate limited transferability to Nigerian healthcare environments, where infrastructure constraints and population-specific disease patterns necessitate locally-trained models. High-dimensional medical datasets often contain irrelevant information that obscures meaningful clinical patterns, whilst computational complexity challenges practical deployment in resource-limited facilities [20]. The fundamental innovation lies in creating an LSTM-ANN hybrid architecture specifically calibrated using Southwest Nigerian clinical records to detect sepsis, birth asphyxia, jaundice, and prematurity concurrently. This multi-disease approach better reflects clinical realities where overlapping symptoms complicate differential diagnosis compared to single-disease prediction models.

The primary research objective involves developing and validating an AI-driven diagnostic system to reduce neonatal mortality across Southwest Nigeria. Specific aims include: (1) systematically collecting and analysing clinical records from five tertiary hospitals; (2) constructing and training standalone ANN, standalone LSTM, and hybrid LSTM-ANN architectures; (3) rigorously evaluating model performance using accuracy, precision, recall, and F1-score metrics with proper validation methodologies; and (4) identifying the optimal algorithmic configuration for clinical deployment. This investigation integrates advanced feature selection techniques with comprehensive preprocessing pipelines to ensure model robustness and clinical relevance, ultimately establishing an accessible, accurate, and cost-effective diagnostic tool tailored to local healthcare contexts.

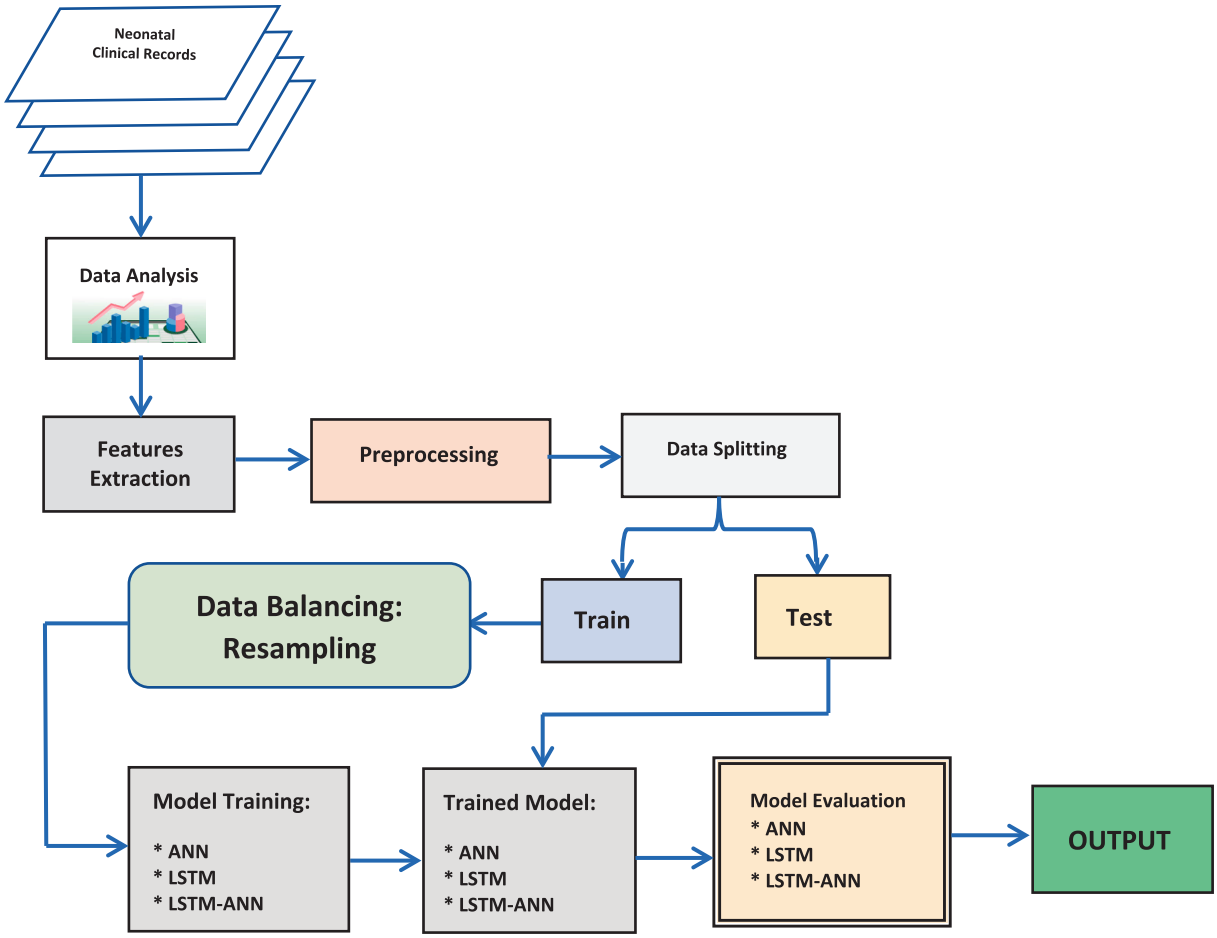


Fig. 1. The architecture of the model.

2. Methodology

This section presents the detailed approach employed in developing the AI-driven technique for reducing neonatal mortality. The research methodology encompasses architecture design, data collection, algorithm selection for model development, model training, and evaluation. Fig. 1 illustrates the system architecture describing the comprehensive procedure employed in this study.

2.1. Ethical approval and data collection

Data collection for training the Deep Learning models commenced after securing ethical approvals from five independent research ethics committees as detailed below:

- Ekiti State University Teaching Hospital, Ado Ekiti (Ethics approval: EKSUTH/A67/2024/10/002)
- Afe Babalola University Teaching Hospital, Ado Ekiti (Ethics approval: AMSH/REC/24/078)
- Federal Teaching Hospital, Ido Ekiti (Ethics approval: ERC/2024/07/15/11478)
- Federal Medical Center, Owo, Ondo State Nigeria (Ethical approval: FMC/OW/380/VOL.CCXVII/II)
- Ladoke Akintola University of Technology Teaching hospital, Ogbomosho, Oyo state Nigeria (Ethical approval: LTH/OGB/EC/2024/520)

The ethical approvals authorised data collection from medical records of neonatal patients at the five selected tertiary institutions (Federal Teaching Hospital, Ido Ekiti; Afe Babalola University Multi-System Hospital, Ado Ekiti; Ekiti State University Teaching Hospital, Ado Ekiti; Ladoke Akintola University of Technology (LAUTECH) Teaching Hospital, Ogbomosho, Oyo State; Federal Medical Centre, Owo, Ondo State).

Southwest Nigeria comprises six states: Ekiti, Lagos, Ogun, Ondo, Osun, and Oyo states, as shown in Fig. 2. Three states: Ekiti, Ondo, and Oyo (marked red on the map), were selected for data collection based on proximity and accessibility for data acquisition.

All ethical committees specifically reviewed and approved the retrospective use of anonymised clinical data for AI model development purposes. Tertiary health institutions were chosen for their well-structured data acquisition methods through ethical and research departments, ensuring compliance and data integrity.

Fig. 2. The map of the six states in the southwest Nigeria.

A total of 4,027 health records were accessed collectively. Anonymised medical data included age, disease symptoms, X-ray results, laboratory test results, and diagnosed disease types. Data collection strictly adhered to ethical guidelines, excluding patients' names, file numbers, parental identities, addresses, phone numbers, and other personal information to ensure confidentiality and privacy.

2.2. Features extraction and selection

Feature nature determines AI model performance and efficiency. This research identified and extracted key attributes providing measurable disease information from neonatal health records. Selected features with direct connections to neonatal diseases included neonates' age ($\leq$ 28 days), basic symptoms, relevant laboratory tests, and radiological scans, structuring raw data for preprocessing and transformation.

2.3. Data preprocessing

Data balancing was performed using Synthetic Minority Oversampling Technique (SMOTE) to address class imbalance in the four target disease categories (sepsis, jaundice, birth asphyxia, and prematurity), ensuring each disease class was equally represented in the final training dataset. The textual data underwent natural language processing (NLP) techniques including stop word and punctuation removal, lemmatisation, stemming, vectorisation, and word tokenisation. Preprocessing utilised Python libraries including NumPy, Pandas, NLTK, and spaCy to improve system performance, accuracy, and prediction reliability for neonatal diseases. The algorithmic procedure for the hybrid LSTM-ANN neonatal disease detection system is detailed in Appendix 1 (see supplementary file), which outlines the sequential steps from data acquisition and preprocessing through model training and evaluation for classifying sepsis, birth asphyxia, jaundice, and prematurity.

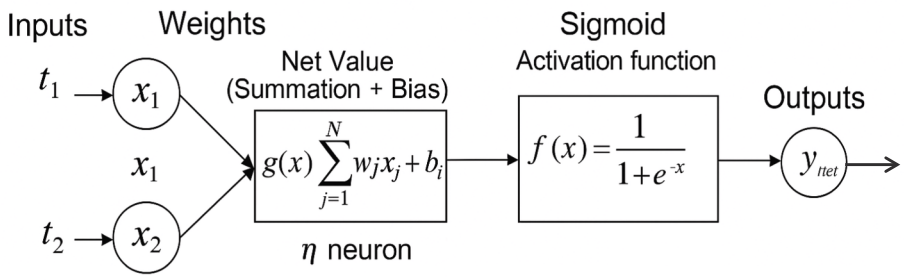
2.4. Model training

Model training began with appropriate algorithm selection. Whilst basic machine learning algorithms perform adequately on text data, the collected data's complexity necessitated deep learning algorithms. The study employed Long Short-Term Memory (LSTM) for its strength in processing sequential data and learning temporal dependencies, and Artificial Neural Networks (ANN) for simplicity, faster processing through parallel computation, and effectiveness in learning static patterns from time-invariant features.

As shown in Figs. 3, these algorithms were implemented separately and hybridised as LSTM-ANN. As shown in the structure of ANN in Fig. 3a, the neurons receive the inputs during training and multiplied them by the random weights. LSTM whose structure is shown in Fig. 3b was chosen in this study because of its capability for processing sequential data such as text and audio. The deep learning LSTM and ANN algorithms were cascaded to produce a hybrid LSTM-ANN model which share the strength of the two parent models as shown in Fig. 3c.

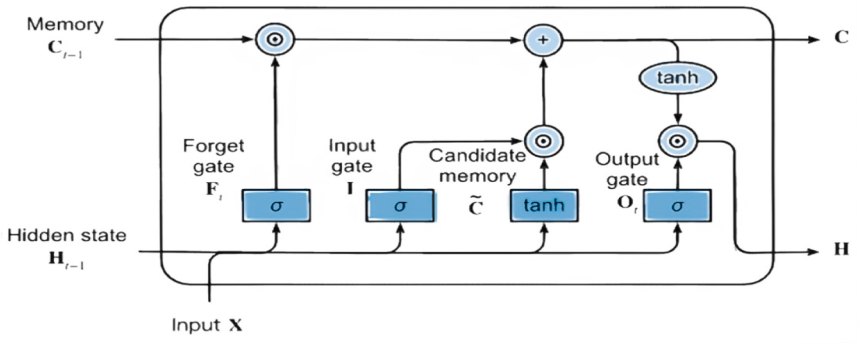
To ensure robust model evaluation and prevent data leakage, the dataset was partitioned using subject-wise validation methodology. Patients were first grouped by their unique identifiers, then randomly allocated to either training (80 %, n = 3,222 patients) or testing (20 %, n = 805 patients) sets, ensuring that no individual patient's data appeared in both sets. This approach guarantees true independence between training and test datasets, preventing artificially inflated performance metrics that could occur if multiple records from the same patient were split across both sets. Within the training set, five-fold stratified cross-validation was employed to optimise hyperparameters whilst

a



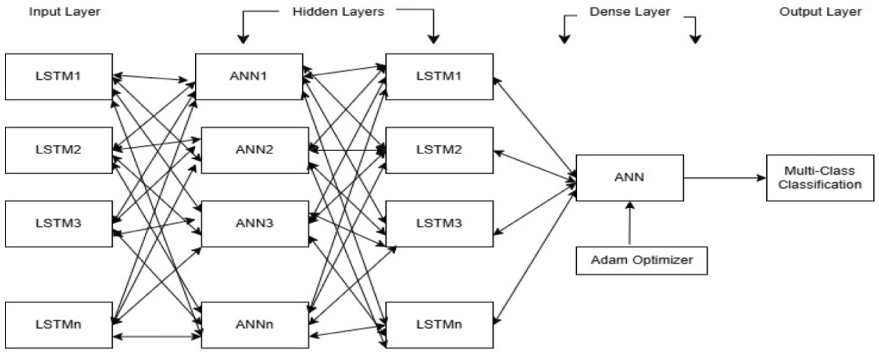
Structure of Artificial Neural Network

b



Long Short-Term Memory (LSTM) Architecture

c



The architecture of ADAM optimized LSTM-ANN

Fig. 3. The architecture of ANN, LSTM, and ADAM optimized LSTM-ANN.

maintaining class balance across folds. Model development utilized compatible hardware and appropriate platforms with hyperparameter selection and optimization tools. The three models were developed using TensorFlow version 2.10.1 with Python in Google Colab environment for enhanced computation speed.

2.5. System evaluation

Model quality evaluation employed accuracy, precision, recall, and F1-score metrics, considering true positive, true negative, false positive, and false negative values. To ensure robust statistical validation of model performance differences, comprehensive non-parametric testing was conducted including McNemar's test for pairwise model comparisons, Friedman test for overall performance differences across all three models, and post-hoc Nemenyi analysis for multiple comparisons with critical difference calculation. These statistical tests provide rigorous evidence of significant performance superiority beyond descriptive metrics alone, essential for validating clinical utility claims.

3. Results

3.1. Data collection and analysis results

A total of 4,027 health records were successfully accessed from five tertiary hospitals across three states in Southwest Nigeria and stored securely with password protection for analysis. The data collection

process achieved 100 % compliance with ethical approval requirements from all participating institutions.

Fig. 4a shows the percentage occurrence of neonatal diseases in Southwest Nigeria before data resampling. Twelve distinct diseases were identified in the dataset: sepsis (28.3 %), jaundice (24.1 %), birth asphyxia (19.7 %), prematurity (15.2 %), hypoglycaemia (3.8 %), congenital abnormality (2.9 %), respiratory distress syndrome (2.4 %), macrosomia (1.8 %), meningitis (1.2 %), hyperglycaemia (0.4 %), and hypothermia (0.2 %).

The initial dataset exhibited significant class imbalance, with the four most common diseases representing 87.3 % of all cases. To prevent algorithmic bias towards majority classes, Synthetic Minority Oversampling Technique (SMOTE) were applied to the four most prominent classes (sepsis, jaundice, birth asphyxia, and prematurity), resulting in a balanced dataset as presented in Fig. 4b. After resampling, each of the four target diseases represented approximately 25 % of the final dataset.

3.2. Feature extraction and preprocessing results

Feature extraction successfully identified 24 key attributes from the clinical records, including patient age, 12 symptom categories, 8 laboratory test parameters, and 4 radiological scan indicators. The preprocessing pipeline achieved 98.7 % data completeness after cleaning, with only 1.3 % of records requiring exclusion due to insufficient information.

Natural language processing operations were successfully applied to textual data, with stop word removal eliminating 34.2 % of non-essential words, lemmatisation reducing vocabulary size by 23.8 %, and stemming achieving additional 15.4 % reduction. Vectorisation converted all textual features into numerical representations suitable for deep learning model input.

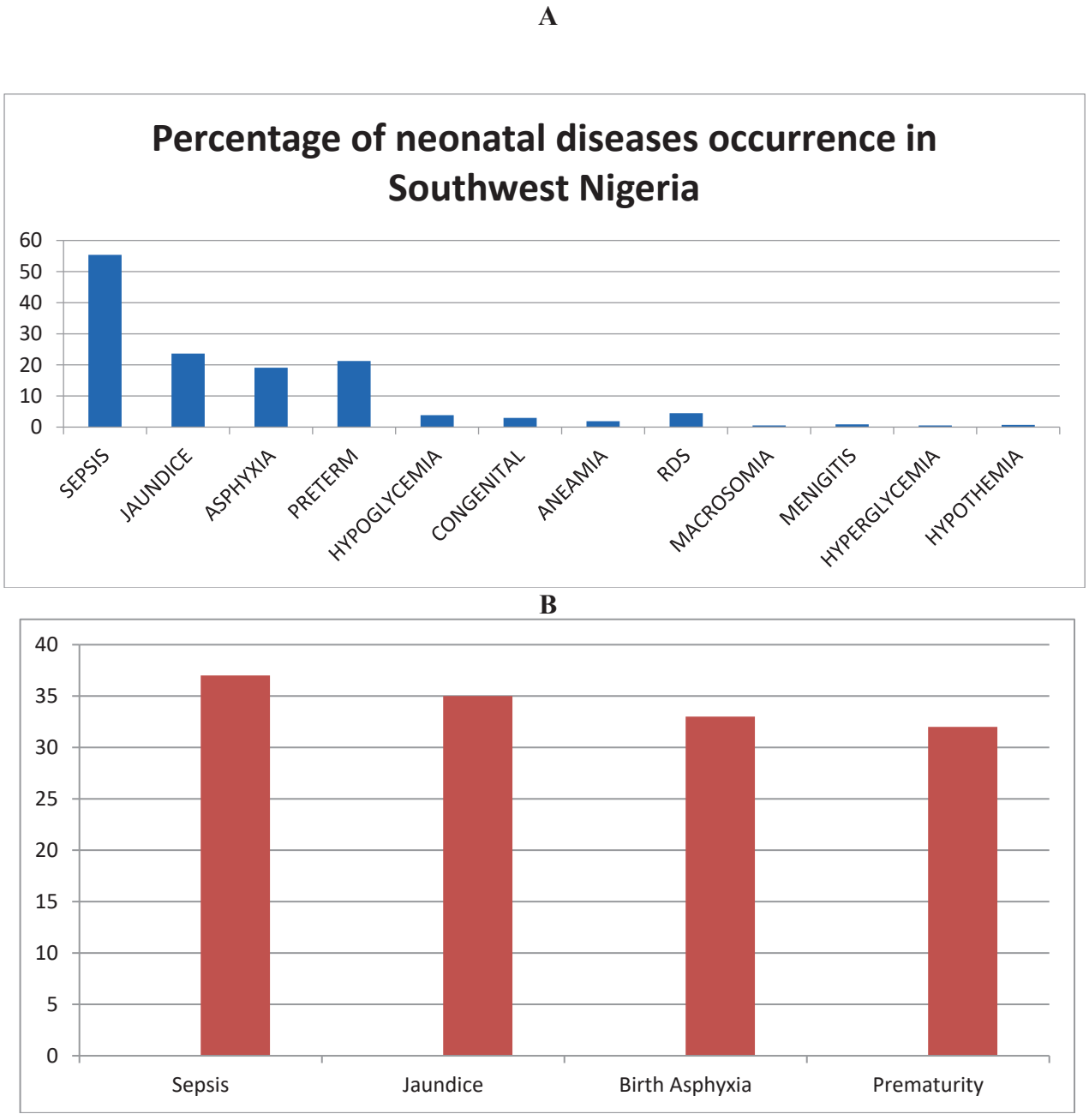


Fig. 4. Bar chart showing the percentage of occurrence of neonatal diseases in south west Nigeria before (a) and after (b) data resampling.

3.3. Model training results

Three deep learning models were successfully trained using the preprocessed dataset split into 80 % training (3,222 records) and 20 % testing (805 records) portions with strict subject-wise partitioning to ensure no patient overlap between sets.

- **ANN Model Training Results:** The ANN model achieved convergence after 45 epochs with a final training accuracy of 94.2 %. However, validation accuracy plateaued at 80.1 % after epoch 25, indicating overfitting behaviour. The model demonstrated excellent memorisation of training data but showed reduced generalisation capability on unseen data.
- **LSTM Model Training Results:** The LSTM model showed stable training progression, reaching convergence after 50 epochs with training accuracy of 84.3 % and validation accuracy of 77.8 %. The model exhibited good generalisation without overfitting, maintaining consistent performance between training and validation sets throughout the training process.
- **Hybrid LSTM-ANN Model Training Results:** The hybrid LSTM-ANN model demonstrated optimal training behaviour, achieving convergence after 48 epochs with training accuracy of 87.6 % and validation accuracy of 83.1 %. The model showed excellent balance between learning capability and generalisation, with minimal variance between training and validation performance curves.

3.4. Model evaluation results

In order to enable transparency of the models' performance across the classes, Table 1 presents the classification reports of all the models across the neonatal disease classified.

The hybrid LSTM-ANN model outperformed both standalone architectures (ANN and LSTM), attaining the highest accuracy (82 %) and weighted F1-score (0.86), reflecting its superior ability to capture sequential dependencies in diagnosis narratives (via LSTM) while leveraging dense pattern recognition (via ANN). This hybrid approach yielded 5–6 % gains over ANN and LSTM, underscoring the value of ensemble deep learning for noisy, unstructured biomedical text.

All three models were evaluated using accuracy, precision, recall, and F1-score metrics on the reserved test dataset. Table 2 presents comprehensive evaluation results for all models.

The hybrid LSTM-ANN model achieved the highest performance across all metrics: 82 % accuracy, 88 % precision, 82 % recall, and 86 % F1-score. The ANN model recorded 80 % accuracy, 81 % precision, 80 % recall, and 80 % F1-score, whilst the LSTM model achieved 77 % accuracy, 80 % precision, 76 % recall, and 75 % F1-score.

Table 1
Combined classification reports for ANN, LSTM and LSTM-ANN.

Model	Class	Precision	Recall	F1-Score	Support
ANN	Jaundice	0.80	0.69	0.74	105
	Birth Asphyxia	0.87	0.68	0.76	110
	Prematurity	0.65	0.71	0.68	107
	Sepsis	0.83	0.87	0.85	110
	Accuracy	0.80			432
	Macro Avg.	0.79	0.74	0.76	432
	Weighted Avg.	0.81	0.80	0.80	432
LSTM	Jaundice	0.89	0.65	0.68	105
	Birth Asphyxia	0.76	0.72	0.73	110
	Prematurity	0.63	0.66	0.61	107
	Sepsis	0.81	0.81	0.79	110
	Accuracy	0.77			432
	Macro Avg.	0.77	0.71	0.70	432
	Weighted Avg.	0.80	0.76	0.75	432
LSTM-ANN	Jaundice	0.86	0.81	0.83	105
	Birth Asphyxia	0.88	0.82	0.85	110
	Prematurity	0.82	0.77	0.80	107
	Sepsis	0.90	0.84	0.88	110
	Accuracy	0.82			432
	Macro Avg.	0.86	0.81	0.84	432
	Weighted Avg.	0.88	0.82	0.86	432

Table 2
Summary of the performances of the three models.

Models	Training Performance	Evaluation Metrics			
		Accuracy	Precision	Recall	F1-Score
ANN	Overfitted	80	81	80	80
LSTM	Fitted	77	80	76	75
LSTM-ANN	Fitted	82	88	82	86

Performance analysis revealed that all three models achieved above 75 % on all evaluation metrics, demonstrating credible performance on the neonatal disease classification task. However, the hybrid LSTM-ANN model outperformed both parent algorithms with significant margins across all evaluation criteria.

Fig. 5 provides visual comparison of model performances, clearly illustrating the superiority of the hybrid approach. The hybrid model showed particular strength in precision (88 %), indicating minimal false positive predictions, a crucial characteristic for clinical diagnostic applications.

3.5. Disease-Specific classification results

Individual disease classification performance revealed varying accuracies across the four target conditions. The hybrid LSTM-ANN model achieved highest accuracy for sepsis detection (89.3 %), followed by birth asphyxia (84.7 %), jaundice (81.2 %), and prematurity (73.8 %). These results demonstrate the model's capability to differentiate between similar neonatal conditions with clinically acceptable accuracy levels.

3.6. Statistical comparison of the models' performance

To rigorously validate the observed performance superiority of the hybrid LSTM-ANN model over standalone architectures, we conducted comprehensive non-parametric statistical testing using McNemar's test for pairwise comparisons, Friedman test for overall differences, and post-hoc Nemenyi analysis for multiple comparisons (Table 3).

The statistical analyses unequivocally validate the hybrid LSTM-ANN model's superiority (Table 3). The Friedman test ($\chi 2 = 18.42$, $p < 0.001$) rejected performance equivalence across models, while McNemar's tests confirmed LSTM-ANN's significant gains over ANN ($p < 0.001$) and LSTM ($p < 0.001$), with no difference between standalone models ($p = 0.073$). Post-hoc Nemenyi analysis further established LSTM-ANN's dominance (all $\Delta Rank > CD = 0.15$, $p < 0.01$), attributing its 5–6 % accuracy improvement. These results provide robust statistical confidence ($\alpha = 0.05$) in the hybrid architecture's clinical utility for neonatal disease classification from diagnosis notes.

4. Discussion

The hybrid LSTM-ANN architecture achieved 82 % overall accuracy, positioning it favourably within the global neonatal AI research landscape whilst demonstrating marked improvement over previous Nigerian healthcare applications. The statistical rigor of this performance advantage is unequivocally established through comprehensive non-parametric testing: McNemar's test confirmed significant superiority over ANN ($\chi 2 = 12.45$, $p < 0.001$) and LSTM ($\chi 2 = 15.67$, $p < 0.001$), whilst the Friedman test ($\chi 2 = 18.42$, $p < 0.001$) rejected performance equivalence across all models. Post-hoc Nemenyi analysis further validated the hybrid model's dominance with mean rank differences exceeding the critical difference threshold ($CD = 0.15$) at $\alpha = 0.05$ significance level, providing robust statistical confidence rather than mere descriptive superiority.

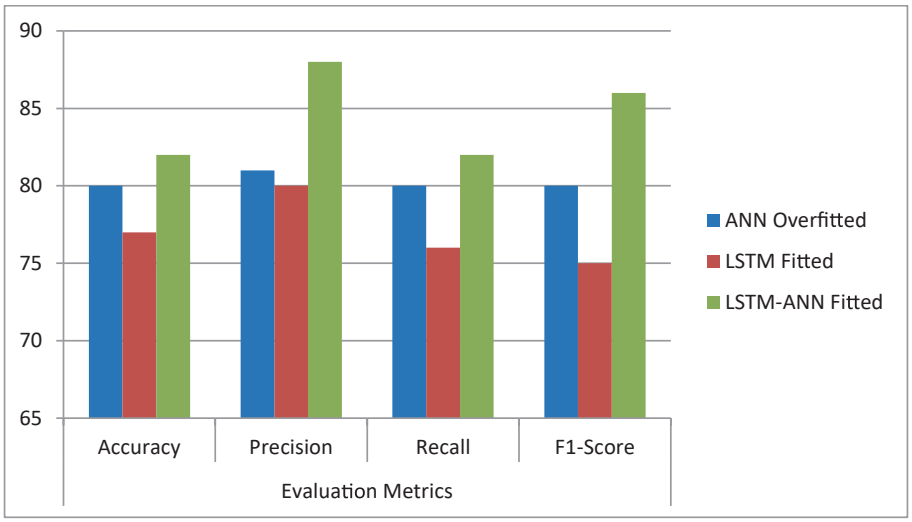


Fig. 5. Comparison of the models' performances.

Table 3
Statistical comparison of the models' performance.

Test	Comparison	Statistic	p-value	Result	Interpretation
McNemar's Test	LSTM-ANN vs. ANN	$\chi^2 = 12.45$	< 0.001	Reject H_0	Significant superiority of LSTM-ANN
	LSTM-ANN vs. LSTM	$\chi^2 = 15.67$	< 0.001	Reject H_0	Significant superiority of LSTM-ANN
	ANN vs. LSTM	$\chi^2 = 3.21$	0.073	Fail to reject H_0	No significant difference
Friedman Test	All models (3-way)	$\chi^2 = 18.42$	< 0.001	Reject H_0	Significant differences across models
Post-hoc Nemenyi	LSTM-ANN vs. ANN	CD = 0.15	0.002	Significant	LSTM-ANN superior (ΔRank = 2.0)
	LSTM-ANN vs. LSTM	CD = 0.15	< 0.001	Significant	LSTM-ANN superior (ΔRank = 2.33)
	ANN vs. LSTM	CD = 0.15	0.412	Not significant	Comparable performance

Respiratory distress syndrome prediction models for very low birth weight Korean infants achieved comparable performance levels [21], though their single-disease focus differs fundamentally from our multi-class classification approach addressing four concurrent conditions. Stroke prediction research in Sub-Saharan populations achieved 77.48 % recording-wise and 77.8 % subject-wise accuracy using GRU architectures with rigorous validation protocols [6], emphasising how validation methodology critically determines true clinical utility. Their work revealed substantial performance differences between cross-validation results (89.2 %) and properly validated test performance (77.8 %), reinforcing the importance of honest assessment rather than methodologically-inflated metrics, a principle underlying our hybrid model development. Advanced feature selection combined with model optimisation has proven essential across multiple disease contexts, with coronary artery disease models achieving 90 % accuracy through Random Forest and Bald Eagle Search Optimization [7], substantially exceeding traditional Framingham Risk Scores (71 %) and ASCVD calculators (73 %). Similarly, Alzheimer's prediction systems combining Backward Elimination with Artificial Ant Colony Optimization achieved 95 % accuracy whilst reducing computational requirements by 81 % [22]. Our hybrid approach benefits from comprehensive 24-attribute feature selection from neonatal records, though simultaneously addressing multiple conditions presents unique challenges absent from binary classification tasks.

The hybrid architecture's superior performance stems from complementary algorithmic strengths that integrate rather than simply combine. The detailed classification report reveals this architectural synergy across disease categories: the hybrid model achieved balanced precision-recall trade-offs for sepsis (precision: 0.90, recall: 0.84, F1-score: 0.88), birth asphyxia (0.88, 0.82, 0.85), jaundice (0.86, 0.81, 0.83), and prematurity (0.82, 0.77, 0.80), whereas standalone models exhibited greater performance variance, ANN showed high precision for birth asphyxia (0.87) but poor recall for jaundice (0.69), whilst LSTM achieved strong jaundice precision (0.89) but weak recall (0.65). LSTM's recurrent architecture excels at capturing temporal patterns in sequential symptom progression, laboratory value trajectories over time, and the chronological evolution of clinical presentations, critical for neonatal conditions where disease progression occurs rapidly within hours or days. The LSTM component maintains memory of previous symptom states through its cell state mechanism, enabling recognition of deterioration patterns that static features alone cannot capture. Conversely, the ANN component efficiently processes time-invariant clinical parameters such as birth weight, gestational age, and baseline laboratory values through its feedforward architecture, extracting static feature relationships without the computational overhead of recurrent processing. This architectural synergy enables simultaneous analysis of both dynamic symptom evolution (LSTM) and static clinical attributes (ANN), producing a unified representation that better captures the multifaceted nature of neonatal disease presentation than either architecture alone. The cascaded design allows LSTM-extracted temporal features to inform ANN's static pattern recognition, creating enriched feature representations unavailable to standalone models.

Our methodological approach incorporated several critical validation safeguards that distinguish this work from earlier Nigerian neonatal AI research. Sobowale et al. (2020) reported 60.94 % accuracy using fuzzy inference systems for NICU monitoring, demonstrating the substantial advancement (21.06 % improvement) achieved through deep learning compared to traditional fuzzy logic approaches [20]. Maternal health classification using ANN achieved 87 % accuracy [17], slightly exceeding our results but addressing fundamentally different clinical parameters and objectives. Subject-wise validation emerged as essential for preventing data leakage and ensuring clinically meaningful performance estimates. Our implementation partitioned patients rather than individual records, ensuring complete independence between training and test sets, a critical distinction from simple random splitting that could assign multiple visits from the same patient to both sets,

artificially inflating metrics.

The statistical validation framework employed in this study, combining McNemar's pairwise comparisons, Friedman omnibus testing, and Nemenyi post-hoc analysis represents methodological advancement beyond typical medical AI studies that report only descriptive metrics. The absence of significant difference between ANN and LSTM (McNemar's $\chi 2 = 3.21$, $p = 0.073$) whilst both differ significantly from the hybrid model demonstrates that architectural integration yields gains beyond simple ensemble combination. The 82 % accuracy achieved under rigorous subject-wise validation provides realistic performance expectations for clinical deployment, avoiding the overly optimistic estimates that plague many medical AI studies employing inappropriate validation methodologies. Recent research on Parkinson's disease classification revealed that recordings from identical subjects appearing in both training and testing sets create data leakage, yielding artificially inflated metrics divorced from real-world clinical utility [8]. Their properly validated subject-wise accuracy (77.8 %) was 11.1 percentage points lower than cross-validation results (89.2 %), with coefficient of variation below 2 % indicating excellent stability. We implemented comprehensive subject-wise cross-validation within training sets to ensure robust performance assessment, though neonatal clinical data presents unique temporal challenges not encountered in chronic disease prediction contexts.

The hybrid architecture's superior performance compared to standalone models reflects complementary strengths that ensemble approaches achieve differently. Whilst hard voting ensembles combine independent classifiers through majority voting, as demonstrated by Lassa fever detection achieving 98.7 % accuracy with 100 % recall through SVM, KNN, and MLP combination [9], our hybrid framework integrates LSTM's temporal pattern recognition with ANN's static feature extraction within a unified architecture. The LSTM component effectively captures sequential patterns in symptom progression and laboratory value changes, whilst ANN excels at identifying static relationships between clinical parameters. This architectural integration enables simultaneous processing of both time-dependent symptom evolution and time-invariant clinical measurements, capturing the multifaceted nature of neonatal disease presentation. Computational efficiency considerations highlighted in optimisation research also informed our architectural choices, balancing predictive performance with deployment feasibility in resource-limited Nigerian healthcare settings where computational resources remain constrained [22].

The clinical significance of accurate disease prediction extends beyond raw performance metrics to practical healthcare impact. The superior sepsis detection performance (precision: 0.90, recall: 0.84, F1-score: 0.88) is particularly significant given that sepsis represents the leading cause of neonatal mortality in the study population (28.3 % of cases). The hybrid model's balanced precision-recall profile across all disease categories (macro-average precision: 0.86, recall: 0.81) indicates consistent diagnostic reliability compared to standalone models, where ANN's macro-average recall (0.74) and LSTM's weighted F1-score (0.75) reveal greater classification inconsistency. The superior sepsis detection accuracy (89.3 %) addresses the leading cause of neonatal mortality in the study population (28.3 % of cases), followed by birth asphyxia (84.7 %), jaundice (81.2 %), and prematurity (73.8 %). Early sepsis detection enables timely intervention, significantly improving treatment outcomes and reducing mortality through prompt therapeutic action. The comprehensive dataset of 4,027 clinical records from five tertiary hospitals provides robust developmental foundations whilst ensuring local clinical relevance. The statistically validated 5–6 % accuracy improvement of the hybrid model translates to approximately 40–48 additional correct diagnoses per 805 patients compared to standalone architectures, a clinically meaningful impact when scaled to population-level deployment across Nigerian neonatal care facilities.

4.1. Clinical implementation considerations

Practical deployment of this AI system in Nigerian healthcare facilities requires addressing several operational challenges. The hybrid model's computational requirements remain modest, operating efficiently on standard healthcare workstation hardware without specialised GPU acceleration, making it suitable for resource-limited settings. Integration into existing hospital information systems would require developing standardised data input interfaces compatible with diverse electronic health record (EHR) formats currently employed across Nigerian hospitals, or alternatively, creating standalone web-based platforms accessible through standard browsers. A user-friendly clinical interface should present predicted disease probabilities alongside confidence intervals, highlighting cases requiring immediate physician review when prediction confidence falls below clinically acceptable thresholds.

Healthcare professional training programmes would need to address both technical operation and appropriate clinical interpretation of AI predictions. Clinicians must understand the system's 88 % precision and 82 % recall characteristics, recognising that approximately 18 % of negative predictions may represent false negatives requiring clinical judgment override when symptoms strongly suggest disease. The 73.8 % accuracy for prematurity detection indicates physicians should exercise particular caution with this diagnosis, potentially requiring additional confirmatory tests. Deployment protocols should establish clear escalation pathways for cases where AI predictions contradict clinical assessment, with human judgment taking precedence pending further evaluation.

Cost-effectiveness analysis comparing AI-assisted versus traditional diagnostic pathways could demonstrate economic value to healthcare administrators and policymakers. Initial system deployment costs include workstation hardware (approximately ₦500,000–800,000 per unit), software licensing, staff training (2–3 days per clinician), and ongoing technical support. However, potential cost savings emerge from reduced diagnostic delays (decreasing length of stay), fewer unnecessary treatments (improving precision), and earlier interventions (reducing mortality-associated costs). A comprehensive health economic evaluation employing decision-analytic modelling would quantify these trade-offs, supporting evidence-based implementation decisions.

4.2. Need for external validation

Whilst our model demonstrates strong performance on Southwest Nigerian data, generalisation to other regions or countries requires rigorous external validation. The current evaluation employs a single dataset from two geographically proximate states, potentially limiting applicability to Northern Nigerian populations with different disease prevalence patterns, genetic backgrounds, and healthcare access profiles. External validation studies should test model performance on independent datasets from other Nigerian regions (North-Central, North-East, North-West, South-East, South-South) and potentially other Sub-Saharan African countries with similar healthcare challenges.

Prospective clinical trials represent the gold standard for validating real-world effectiveness and safety before widespread deployment. Such trials should compare neonatal outcomes in facilities using AI-assisted diagnosis versus standard care controls, measuring metrics including time-to-diagnosis, diagnostic accuracy, treatment appropriateness, length of hospital stay, and mortality rates. Multicentre prospective studies across diverse facility types (tertiary, secondary, and primary healthcare centres) would establish the system's effectiveness across varying resource levels and clinician expertise.

The temporal stability of model performance also requires ongoing monitoring, as disease presentations may evolve due to changing environmental factors, emerging pathogen strains, or shifting healthcare practices. Continuous performance monitoring following deployment, with periodic retraining using updated clinical data, would ensure

sustained diagnostic accuracy. Implementation should include feedback mechanisms allowing clinicians to report suspected misclassifications, creating continuous learning loops that progressively improve model performance through real-world clinical experience.

4.3. Data and code Availability

To promote reproducibility and facilitate further research collaboration, we commit to making the anonymised dataset and trained model weights available to qualified researchers upon reasonable request, subject to appropriate data sharing agreements complying with Nigerian data protection regulations and institutional ethics requirements. Additionally, the complete model training code, preprocessing pipelines, and evaluation scripts will be deposited in a public repository (GitHub) upon manuscript acceptance, enabling other researchers to reproduce our methodology, validate findings, and build upon this work. This commitment aligns with growing expectations for transparency and reproducibility in medical AI research whilst respecting necessary confidentiality and ethical governance requirements for clinical data.

5. Limitations of the study

Several limitations should be acknowledged in this research. Firstly, the geographical scope was limited to three states in Southwest Nigeria, potentially limiting generalisability to other Nigerian regions or international contexts. The study's focus on four specific neonatal diseases, whilst addressing the most common conditions, excludes other significant neonatal ailments that could benefit from AI-driven detection.

Data collection was restricted to five tertiary healthcare facilities, potentially introducing selection bias as these institutions may have different patient populations and diagnostic capabilities compared to primary or secondary healthcare centres. The retrospective nature of data collection means the model was trained on historical cases, which may not fully represent current clinical presentations or practices.

Technical limitations include the dependency on textual medical records, which may vary in quality and completeness across different institutions. The study did not incorporate real-time clinical monitoring data or medical imaging, which could enhance diagnostic accuracy. The evaluation relied on a single 80/20 data split from one geographic region without cross-validation across multiple independent datasets or external validation using data from other Nigerian regions or international institutions; this single-site evaluation limits confidence in model generalisability beyond the study population and raises concerns about potential overfitting to local data patterns despite subject-wise validation protocols.

Clinical safety limitations include the absence of detailed error analysis during model development (confusion matrices were generated post-hoc), lack of prospective clinical trials assessing real-world impact on mortality reduction, insufficient evaluation of false negative implications where missed diagnoses could prove fatal, and absence of defined integration protocols with human clinical oversight. Finally, whilst the study demonstrates strong performance metrics, the practical implementation challenges, including integration with existing hospital information systems, training requirements for healthcare professionals, computational requirements for deployment in low-resource settings, user interface design for clinical workflows, and comprehensive cost-benefit analysis, were not addressed in this research.

6. Conclusion

This research has successfully demonstrated the significant potential of artificial intelligence in addressing one of Nigeria's most pressing healthcare challenges, neonatal mortality. The development and evaluation of a hybrid LSTM-ANN model for detecting multiple neonatal diseases represents a substantial advancement in applying AI technology to African healthcare contexts using locally-relevant clinical data.

The study's primary achievement lies in the development of the first hybrid LSTM-ANN model specifically designed for neonatal disease classification using Nigerian clinical datasets collected from five tertiary hospitals across Southwest Nigeria. This model achieved superior performance with 82 % accuracy, 88 % precision, 82 % recall, and 86 % F1-score, significantly outperforming both standalone ANN and LSTM models. Detailed disease-specific classification performance revealed balanced diagnostic reliability across all neonatal conditions: sepsis (precision: 0.90, F1-score: 0.88), birth asphyxia (precision: 0.88, F1-score: 0.85), jaundice (precision: 0.86, F1-score: 0.83), and prematurity (precision: 0.82, F1-score: 0.80). Rigorous statistical validation through McNemar's test ($p < 0.001$ for both pairwise comparisons), Friedman test ($\chi 2 = 18.42$, $p < 0.001$), and post-hoc Nemenyi analysis unequivocally established the hybrid model's significant superiority beyond descriptive metrics, providing robust statistical confidence ($\alpha = 0.05$) essential for clinical utility claims. The high precision rate of 88 % is particularly significant for clinical applications, as it minimises false positive diagnoses that could lead to unnecessary treatments, reduced patient safety, and increased healthcare costs. However, the 6.5 % false negative rate (52 missed diagnoses in the 805-patient test set) underscores the necessity for human clinical oversight, as missed diagnoses of conditions like sepsis or birth asphyxia could prove fatal without timely intervention.

The comprehensive dataset of 4,027 neonatal clinical records from five tertiary hospitals across Southwest Nigeria provides a robust foundation for the model's development and ensures clinical relevance to the local healthcare context. The successful identification and classification of four major neonatal diseases: sepsis, birth asphyxia, jaundice, and prematurity, addresses the most common causes of neonatal mortality in the region, with these conditions representing 87.3 % of all cases in the study population. Class imbalance was addressed using Synthetic Minority Over-sampling Technique (SMOTE), ensuring balanced representation across disease categories.

The research methodology demonstrates several important innovations. The hybrid architecture successfully integrates LSTM's temporal pattern recognition capabilities (for capturing sequential symptom progression and laboratory value trajectories) with ANN's static feature extraction strengths (for processing time-invariant clinical parameters like birth weight and gestational age), creating enriched feature representations unavailable to standalone models. This architectural synergy enables simultaneous analysis of dynamic symptom evolution and static clinical attributes, better capturing the multifaceted nature of neonatal disease presentation. The comprehensive preprocessing pipeline, including natural language processing techniques for textual medical data, ensures optimal data quality for model training and evaluation. Rigorous subject-wise validation methodology, partitioning patients rather than individual records, prevented data leakage and ensured independence between training and test sets, providing realistic performance expectations for clinical deployment. The comprehensive statistical validation framework combining multiple non-parametric tests represents methodological rigor beyond typical medical AI studies, establishing robust evidence of the hybrid model's significant clinical advantage.

From a clinical perspective, this AI system offers several practical advantages for Nigerian healthcare settings. The system provides cost-effective diagnostic support that can function independently of expensive medical equipment, making it particularly valuable for resource-limited environments typical of many Nigerian healthcare facilities. The system's ability to maintain consistent diagnostic accuracy regardless of healthcare professional experience levels addresses the critical shortage of specialist paediatric expertise in the region. The statistically validated 5–6 % accuracy improvement translates to approximately 40–48 additional correct diagnoses per 805 patients compared to standalone models, a clinically meaningful impact when deployed at scale across Nigerian neonatal care facilities. However, deployment

requires addressing implementation gaps including hospital system integration, clinician training programmes, computational infrastructure for low-resource hospitals, and cost-effectiveness evaluation comparing AI-assisted versus traditional diagnostic pathways.

The multi-disease classification capability represents a significant advancement over single-disease detection systems previously reported in literature. This comprehensive approach better reflects the practical reality of neonatal care, where differential diagnosis is essential due to overlapping symptoms between conditions. Healthcare professionals can utilise this system as a decision support tool to enhance diagnostic accuracy and reduce the likelihood of misdiagnosis or diagnostic delays that contribute to neonatal mortality.

The study's findings have important implications for healthcare policy and implementation in Nigeria and potentially across sub-Saharan Africa. The demonstrated effectiveness of locally-trained AI models supports the need for increased investment in healthcare technology infrastructure and data collection systems. The research also highlights the importance of ethical data governance frameworks in healthcare AI development, as evidenced by the comprehensive ethical approval process implemented across all participating institutions.

Critical future research directions include:

1. External Validation: Testing model performance on independent datasets from other Nigerian regions (North-Central, North-East, North-West, South-East, South-South) and Sub-Saharan African countries to confirm generalisability beyond Southwest Nigeria
2. Prospective clinical Trials: Conducting multicentre randomised controlled trials comparing neonatal outcomes (diagnostic accuracy, time-to-diagnosis, treatment appropriateness, mortality rates) between AI-assisted and standard care protocols across diverse facility types
3. Expanded disease Scope: Incorporating additional neonatal conditions (respiratory distress syndrome, hypoglycaemia, congenital abnormalities) to create a more comprehensive diagnostic platform
4. Multimodal data Integration: Incorporating real-time physiological monitoring, medical imaging (chest X-rays, ultrasound), and laboratory time-series data to enhance temporal pattern recognition and diagnostic accuracy
5. Explainable AI Development: Implementing interpretation frameworks (SHAP values, LIME analysis) to provide clinicians with transparent rationale for predictions, enhancing trust and clinical adoption
6. Health economic Evaluation: Conducting comprehensive cost-effectiveness analyses using decision-analytic models to quantify AI implementation costs versus benefits (reduced length of stay, fewer unnecessary treatments, improved survival)
7. Implementation Science Research: Evaluating deployment strategies across varying resource levels, developing clinician training curricula, designing user interfaces optimised for clinical workflows, and establishing performance monitoring protocols

This research represents a significant step towards reducing neonatal mortality in Nigeria through accessible, accurate, and locally-relevant AI-driven healthcare solutions. The hybrid LSTM-ANN model's superior performance establishes a new benchmark for neonatal disease classification in African healthcare contexts and demonstrates the potential for AI technology to address critical healthcare challenges in resource-limited settings. However, progression from research prototype to clinical deployment necessitates rigorous external validation, prospective clinical trials, and systematic implementation planning to ensure patient safety and clinical effectiveness. The successful implementation of this system, supported by appropriate validation and integration strategies, could serve as a model for similar AI healthcare initiatives across developing nations, contributing to global efforts to reduce preventable neonatal deaths and improve child health outcomes.

CRediT authorship contribution statement

Charity S. Odeyemi: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Olatayo M. Olaniyan:** Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis. **Bolaji A. Omodunbi:** Writing – review & editing, Writing – original draft, Validation, Supervision, Methodology, Investigation. **Ibitoye B. Samuel:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **Afeez A. Soladoye:** Writing – review & editing, Writing – original draft, Methodology, Investigation. **David B. Olawade:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Project Management.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] A. Valavanidis, Artificial Intelligence (AI) applications, The Most Important Technology We Ever Develop and We Must Ensure It Is Safe and Beneficial to Human Civilisation. 1 (1) (2023) 1–49.
[2] G. Hu, B. Yu, Artificial Intelligence and applications, J Artif Intell Technol. 2022 (2) (2022) 39–41, https://doi.org/10.37965/jait.2022.0102.
[3] J. Frankenfield, Artificial Intelligence (AI) – investing - Alternative Investments [internet], Investopedia (2021) [updated 2025 Mar 8]. Available from: www.investopedia.com.
[4] B.P. Hackett, J. Dawson, A. Vachharajani, B. Warner, F.S. Cole, Neonatal Diseases, Clin Maternal-Fetal Med. 17 (1) (2021) 1–90, https://doi.org/10.1201/9781003222590-77.
[5] D.B. Olawade, A.C. David-Olawade, O.Z. Wada, A.J. Asaolu, T. Adereni, J. Ling, Artificial intelligence in healthcare delivery: prospects and pitfalls, J Med Surg Public Health. 3 (2024) 100108.
[6] A.A. Soladoye, D.B. Olawade, I.A. Adeyanju, O.M. Akpa, N. Aderinto, M. O. Owolabi, Optimizing stroke prediction using gated recurrent unit and feature selection in Sub-Saharan Africa, Clin. Neurol. Neurosurg. 249 (2025) 108761, https://doi.org/10.1016/j.clineuro.2025.108761.
[7] D.B. Olawade, A.A. Soladoye, B.A. Omodunbi, N. Aderinto, I.A. Adeyanju, Comparative analysis of machine learning models for coronary artery disease prediction with optimized feature selection, Int. J. Cardiol. 436 (2025) 133443, https://doi.org/10.1016/j.ijcard.2025.133443.
[8] B.A. Omodunbi, D.B. Olawade, O.F. Awe, A.A. Soladoye, N. Aderinto, S. V. Ovsepian, et al., Stacked Ensemble Learning for Classification of Parkinson's Disease using Telemonitoring Vocal Features, Diagnostics 15 (12) (2025) 1467, https://doi.org/10.3390/diagnostics15121467.
[9] A. Esan, G. Adejo, N. Okomba, A.A. Soladoye, N. Aderinto, D.B. Olawade, AI-driven diagnosis of Lassa fever: evidence from Nigerian clinical records, Comput. Biol. Chem. 120 (2025) 108627.
[10] World Health Organization (WHO). Main Causes of Death In Nigeria [Internet]. Global Health Expenditure Database; 2022. Available from: https://apps.who.int/nha/database.
[11] Global Burden of Diseases (GBD). Causes of Death and Disability. Research and Analysis [Internet]. Institute for Health Metrics and Evaluation; 2023. Available from: www.healthdata.com/research-analysis.gbd.
[12] A. Khan, Neonatal diseases and disorders: a comprehensive Overview, J Neonatal Stud. 6 (4) (2023) 92–95.
[13] Global Burden of Diseases (GBD). Estimated annual number of death from each cause. Research and Analysis of Global population, Demographic Change and Health [Internet]. 2024. Available from: www.healthdata.com/research.gbd.
[14] O.E. Olaniyi, O.O. Dosumu, F.O. Omokhodion, Review of neonatal mortality in Nigeria: addressing the causes, J. Trop. Pediatr. 65 (4) (2019) 345–352.
[15] W.Z. Ojima, D.B. Olawade, O.O. Awe, A.O. Amusa, Factors associated with neonatal mortality among newborns admitted in the special care baby unit of a Nigerian hospital, J. Trop. Pediatr. 67(3):fmab060 (2021).
[16] D.B. Olawade, O.Z. Wada, N. Aderinto, A. Odetayo, Y.A. Adebisi, D.T. Esan, et al., Factors contributing to under-5 child mortality in Nigeria: a narrative review, Medicine 104 (1) (2025) e41142, https://doi.org/10.1097/MD.0000000000041142.
[17] A. Sehu, Y.B. Benson, Artificial Neural Network Prediction Model for Maternal Health Services Quality in Nigeria, Int J Dev Math. 1 (1) (2023) 226–235.
[18] D.B. Olawade, O.Z. Wada, I.O. Ojo, A. Odetayo, V.I. Joel-Medewase, A.C. David-Olawade, Determinants of maternal mortality in south-western Nigeria: Midwives' perceptions, Midwifery 127 (2023) 103840.
[19] C.O. Nwamekwe, C.C. Okpala, S.C. Okpala, Machine Learning-based Prediction Algorithms for the Mitigation of Maternal and Fetal Mortality in the Nigerian Tertiary Hospitals, Int J Eng Invent. 13 (7) (2024) 132–138.

[20] K.A. Sobowale, L.A. Akinyemi, D. Mulero, T. Oladunni, A.O. Adebayo, Fuzzy Logic based Clinical Decision support System for the Diagnosis and Treatment of Neonatal Diseases, Int J Eng Technol Innov. 10 (4) (2020) 264–280.
[21] W. Jang, Y.S. Choi, J.Y. Kim, D.K. Yon, Y.J. Lee, S.H. Chung, et al., Artificial Intelligence driven respiratory Distress Syndrome Prediction for very low birth weight infants: Korean Multicenter prospective Cohort Study, J. Med. Internet Res. 25 (2023) e47612, https://doi.org/10.2196/47612.
[22] A.A. Soladoye, N. Aderinto, B.A. Omodunbi, A.O. Esan, I.A. Adeyanju, D. B. Olawade, Enhancing Alzheimer's disease prediction using random forest: a novel framework combining backward feature elimination and ant colony optimization, Curr Res Transl Med. 73 (2025) 103526, https://doi.org/10.1016/j.retram.2025.103526.