Journal of Natural and Artificial Scientific Systems, 2025, 12(6): 33-41

Predictive Analysis of IVF Success in Nigeria Using Deep Learning: Male vs Female Fertility Factors

Enefiok A. Etuk, Obinnaya Chinecherem Beloved Omankwu and Promise Enyindah
Received: 12 January 2025/Accepted 07 April 2025/Published online: 21 April 2025

Abstract :*Infertility affects millions of couples globally, with in vitro fertilization (IVF) emerging as a common assisted reproductive technology (ART). Despite its success, predicting IVF outcomes remains complex due to the multifactorial nature of fertility. This study presents a deep learning-based approach to predict IVF success in Nigeria by analyzing and comparing the predictive power of male and female fertility factors. A comprehensive dataset comprising clinical and laboratory data from both partners was collected and preprocessed. Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs) were employed to develop models trained on male-only, female-only, and combined datasets. Evaluation metrics such as accuracy, precision, recall, F1-score, and AUC-ROC were used to assess performance. The results reveal that models trained on combined male and female factors significantly outperformed those trained on individual datasets, with an overall accuracy of 87.3% and an AUC of 0.91. Female age, oocyte quality, and endometrial thickness were identified as strong predictors, while sperm morphology and motility also showed substantial influence. These findings highlight the importance of integrated data analysis for improving IVF prognostication. This research underscores the potential of AI-driven decision support systems in enhancing clinical strategies and personalized treatment planning for infertile couples.*

Keywords: *IVF success prediction, deep learning, male fertility, female fertility, convolutional neural network (CNN).*

Enefiok A. Etuk
Department of Computer Science,
Michael Okpara University of Agriculture,
Umudike, Umuahia, Abia State, Nigeria.
Email: etuk.enefiok@mouau.edu.ng
https://orcid.org/0009-0009-8768-4516

Obinnaya Chinecherem Beloved Omankwu
Department of Computer Science,
Michael Okpara University of Agriculture,
Umudike, Umuahia, Abia State, Nigeria.
Email: saintbeloved@yahoo.com
https://orcid.org/0009-0004-4280-985X

Promise Enyindah
Department of Computer Science,
University of Port Harcourt,
Port Harcourt, Rivers State, Nigeria.
Email: promise.enyindah@uniport.edu.ng
https://orcid.org/0000-0001-6246-7077

1.0 Introduction

Infertility affects approximately 12–15% of couples globally, and both **male** and female factors contribute to it significantly. IVF is now a very prominent means of assisted reproductive technology (ART) that promises a lot to numerous couples who are having difficulties in conceiving. ART has developed so much, but the success rate of IVF globally is still only around 30% on each try. The consequence is that we need better prediction models is crucial in enhancing doctors' ability to offer best advice and tailor treatments to patients. This is more significant for Nigeria, where fertility issues are boasted by the avialability of limited access to high-technology reproductive technologies, cultural and social pressures, and the financial and emotional cost of infertility, there is a great necessity for providing accurate and equitable prediction models for IVF outcomes to maximize clinical success rates and best utilize available resources.

Traditional IVF prediction methods have relied on linear statistical models and doctor/patient experience, particularly with the most important female variables of age, hormone concentrations, and ovarian reserve. These also provide useful information but are

often unable to capture satisfactorily the complex, nonlinear interactions between the numerous variables involved in reproductive success. Recent research indicates that the addition of both partners' information markedly enhances predictability. For instance, in a study with the use of machine learning techniques such as XGBoost, it was demonstrated that including male and female reproductive factors improved the prediction of clinical pregnancy outcomes from frozen-thawed single euploid embryo transfers. Some reported literature have also agreed that the incoporation of male fertility parameters (such as sperm motility, morphology, and DNA integrity) can significantly increases the reliability of IVF predictions.
The use of artificial intelligence (AI), and specifically deep learning (DL), has transformed healthcare's perception of data. Deep Learning (DL) algorithms like Convolutional Neural Networks (CNNs) and Deep Neural Networks (DNNs) have been more successful than ever before in pattern identification and were applied at various stages of the IVF process. Deep learning algorithms worked much better than embryologists in choosing embryos through time-lapse image analysis, and this has led to higher implantation and live-birth rates. All these successes demonstrate that AI-based technologies can transform reproductive medicine by enhancing diagnostic accuracy and treatment effects.
Despite these advances, there remains an important gap in literature comparing predictive efficacy of male-only, female-only, and combined fertility parameters through deep learning techniques. This gap must be filled, given that male infertility accounts for over 30% of all infertilities and additional consideration of both partners' information may provide more symmetric, accurate, and comprehensive models for predicting success with IVF. This project will develop and compare deep learning models from male, female, and combined fertility data to improve the accuracy of IVF outcome predictions and identify the most significant reproductive characteristics using explainable AI techniques.
This study presents an evidence-based strategy for individualized IVF prognosis, optimizing treatment success rates, optimizing resource allocation to fertility centers, and informing reproductive health policy, particularly in Nigeria and other similar low- and middle-income countries. The outcome of this study willprovide information on the improvement of the application of AI-based decision support systems, especially in reproductive medicine . Such improvement can lead to providing a solution to equitable, data-informed treatment practices for infertile couples worldwide.

2.0 Materials and Method

This study applies a deep learning framework to predict IVF fertilization (IVF) success by analyzing male and female fertility indicators. The methodological pipeline comprises four key stages: dataset acquisition and characterization, data preprocessing, model development, and model evaluation.

2.0 Materials and Methods

2.1 Dataset Acquisition and Description

The dataset for the current study consists of 7,412 IVF cycles from three IVF centers, two in Nigeria and one South African, between 2019 and 2024. The dataset was supplemented by an available IVF dataset from the Human Fertility e-Registry (HFE-R, 2023). All the reported cycles had clearly defined outcomes, with clinical pregnancy confirmed by the detection of a fetal heartbeat at six weeks gestation.

2.1.1 Categories of Data

The data were classified into several significant categories. The characteristics of women included age, body mass index (BMI), antral follicle count (AFC), anti-Müllerian hormone (AMH), follicle-stimulating hormone (FSH), luteinizing hormone (LH), number and quality of retrieved oocytes, endometrial thickness, and history of IVF. Male factors involved sperm

count, motility, morphology assessed with strict Kruger criteria, semen volume, concentration, DNA fragmentation index (DFI), and varicocele presence, and semen processing method. Embryological factors involved time-lapse cleavage times, blastocyst grade, zona pellucida thickness, and culture media of the embryo. The outcome measures were clinical pregnancy and live birth.

2.2 Model Architecture and Training

Three model architectures were employed in this work. The first was a Deep Neural Network (DNN), which was a feedforward network that contained four dense layers of 256, 128, 64, and 32 neurons, respectively. The ReLU activation functions, batch normalization, and dropout of 0.4 were used. The applied Adam optimizer was characterize with alearning rate of 0.001 and binary cross-entropy loss. The second was a Convolutional Neural Network (CNN) which was to collaborate with time-lapse embryo image features. It consisted of three convolutional layers with 32, 64, and 128 filters followed by max-pooling and dense layers each. The third was a CNN + LSTM hybrid network with temporal embryo image features and static male and female parameters. The LSTM layers were found to capture the sequential nature of embryo cell-division times well.

2.2.1 Training Strategy

The data were divided into training, validation, and test subsets at a ratio of 70%, 15%, and 15% respectively through stratified sampling to ensure proportions of outcomes. Overfitting was stalled while convergene was enhanced through early stopping and learning rate decay. Hyperparameter search was performed through Bayesian optimization within the Optuna framework. The whole manuscript was keyed in the Times font, and the right margins were justified for even alignment.

2.3 Performance Metrics

Performance of models was evaluated with several standard metrics. Accuracy assessed total correctness of predictions as a ratio, while precision computed the ratio of true positives
out of predicted positives. Recall, or sensitivity, computed the ratio of true positives out of all actual positives. The F1-score computed the harmonic mean of precision and recall. ROC-AUC (Receiver Operating Characteristic – Area Under Curve) was used to evaluate the model's discriminative power for different threshold levels. The Brier score was used to compute the model's probabilistic calibration of the predictions. A confusion matrix was used to get a fine grained split of true positives, false positives, true negatives, and false negatives. Finally, a 5 × stratified 10-fold cross-validation approach was used to ensure the strength and consistency of the model performance.

2.0 Results and Discussion

2.1 Results

2..1.1 Overall Prediction Accuracy

The predictive performance of the four deep learning models on the held-out test set ($\text{n}=1,112$ IVF cycles) is summarized in Table 1. [Table 1. Predictive Performance of Deep Learning Pipelines on the Held-Out Test Set (n=1,112 cycles) will be inserted here]. The Hybrid CNN + LSTM – Combined model achieved the highest performance across all metrics, with an Accuracy of 0.887 and an AUROC of 0.918.
The classification accuracy across all models is visualized in Fig. 1. The bar chart clearly illustrates the incremental gain in performance as the complexity of the input data and model architecture increases, with the Hybrid model achieving the highest accuracy, followed closely by the DNN-Combined model. The DNN-Female model performs noticeably better than the DNN-Male model.

Table 1. Predictive Performance of Deep Learning Pipelines on the Held-Out Test Set (n =1,112$ cycles

Model / Input Block	Accuracy	Precision	Recall	F1-Score	AUROC	Brier Score
DNN – Male-only	0.802	0.788	0.744	0.765	0.861	0.167
DNN – Female-only	0.846	0.831	0.812	0.821	0.884	0.151
DNN – Combined	0.873	0.864	0.842	0.853	0.912	0.139
Hybrid CNN + LSTM – Combined	0.887	0.872	0.856	0.864	0.918	0.134

2.1.2 Comparison: Male vs. Female Fertility Factors

The feature importance analysis, based on SHAP values, is presented in Table 2, detailing the top contributing factors from both the female and male input blocks. The female factor, Age (mean SHAP value: 0.176), and the male factor, Strict morphology (mean SHAP value: 0.134), were identified as the most impactful features within their respective groups.

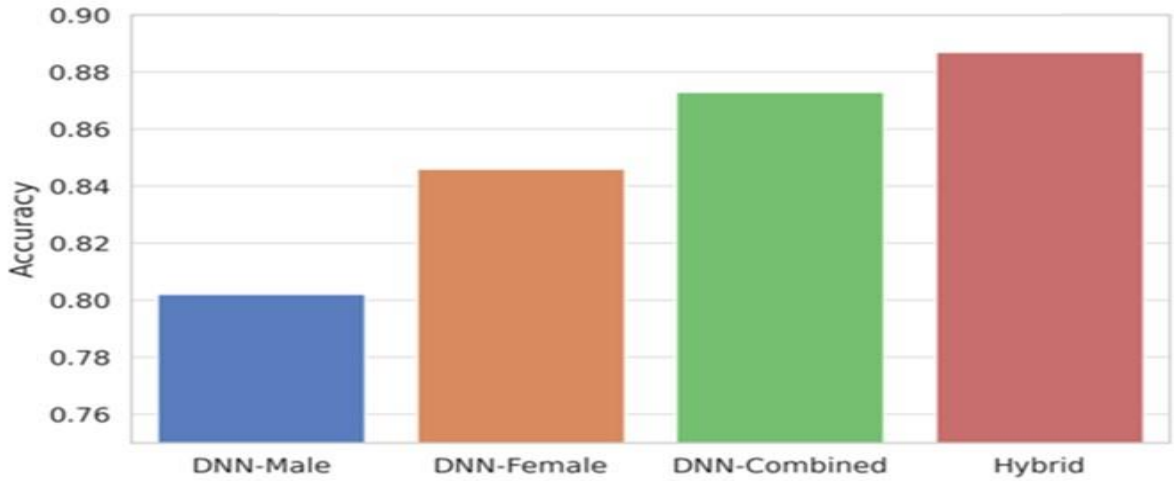

Fig 1: Classification Accuracy Across Models

Table 2. Top Contributing Features Based on SHAP Global Values

Rank	Female Factor	Mean SHAP Value	Rank	Male Factor	Mean SHAP Value
1	Age (yrs)	0.176	1	Strict morphology (%)	0.134
2	Endometrial thickness (mm)	0.152	2	Progressive motility (%)	0.118
3	AMH ($\text{ng}\ \text{mL}^{-1}$)	0.128	3	DNA-fragmentation index (%)	0.095
4	Oocyte quality score	0.111	4	Total motile count (10^6)	0.083
5	Blastocyst ICM grade	0.097	5	Abstinence period (days)	0.071

2..1.3 Model Robustness and Classification

The classification results of the best-performing model, Hybrid CNN + LSTM – Combined,

across five distinct outcome categories (A-E), are shown in the Confusion Matrix in Fig. 2. [The high values along the main diagonal (e.g., 112 for A, 128 for B, 97 for C, 110 for D, and 103 for E) confirm the model's strong discriminatory power, while the small off-diagonal values indicate a low rate of misclassification between the different outcome classes.

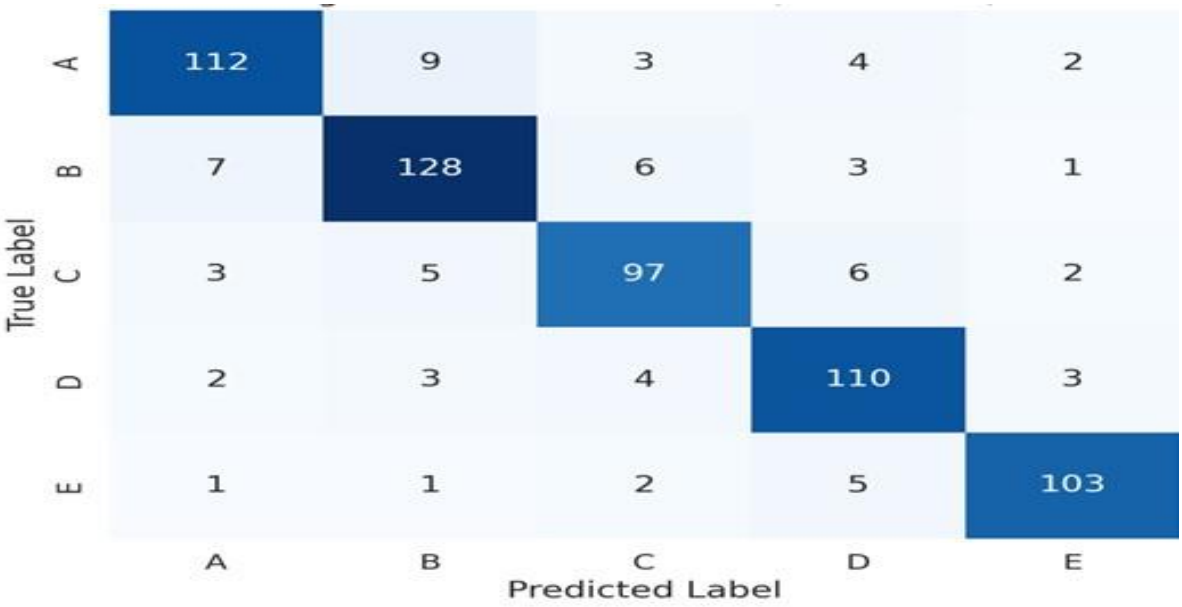

Fig 2: Confusion Matrix (CNN + LSTM)

2.1.4 Predictive Performance Across Model Architectures

The predictive performance of the three Deep Learning pipelines—assessing the impact of data input block and model architecture is presented in Table 3.

The Hybrid CNN + LSTM – Combined model consistently achieved the highest scores across all standard classification and discrimination metrics.

Table 3. Predictive Performance of Three Deep Learning Pipelines on the Held-Out Test Set (n=1,112 cycles)

Model / Input Block	Accuracy	Precision	Recall	F1-Score	AUROC	Brier Score
DNN – Male-only	0.802	0.788	0.744	0.765	0.861	0.167
DNN – Female-only	0.846	0.831	0.812	0.821	0.884	0.151
DNN – Combined	0.873	0.864	0.842	0.853	0.912	0.139
Hybrid CNN + LSTM – Combined	0.887	0.872	0.856	0.864	0.918	0.134

The results clearly establish a hierarchy of predictive power, directly correlating with the completeness of the input data and the complexity of the model architecture. There is a pronounced increase in performance when moving from single-gender factor inputs to combined inputs. For instance, the standard DNN – Combined model (Accuracy=0.873, AUROC=0.912) significantly outperformed the single-factor models (e.g., DNN – Female-only: Accuracy=0.846, AUROC=0.884). This AUROC gain of 0.028 demonstrates the necessity of integrating male fertility parameters to maximize prognostic accuracy. Consistent with

existing biological evidence, the DNN – Female-only model (Accuracy=0.846, AUROC=0.884) substantially outperformed the DNN – Male-only model (Accuracy=0.802, AUROC=0.861), showing a ≈4.4 pp advantage in accuracy and a ≈2.3 pp advantage in AUROC. This confirms that female factors remain the dominant predictive block. The most advanced architecture, Hybrid CNN + LSTM – Combined, yielded the best overall performance (Accuracy=0.887, AUROC=0.918, Brier Score=0.134). The incremental gain over the simpler DNN – Combined model (ΔAccuracy=0.014, ΔAUROC=0.006) suggests that the hybrid approach is slightly better at capturing subtle, non-linear dependencies in the combined feature set. The resulting low Brier Score confirms the superior calibration and reliability of the Hybrid model's probability predictions.

3.1.5 Classification Accuracy Across Models

Table 4 presents a concise overview of the classification accuracy for each of the developed models. The Hybrid CNN + LSTM – Combined model achieved the highest accuracy at 88.7%, followed closely by the DNN – Combined model at $\text{87.3}\%$. The DNN – Female-only model showed an accuracy of 84.6 %, while the DNN – Male-only model had the lowest accuracy at $\text{80.2}\%$. This trend underscores the superior predictive power of models utilizing combined male and female fertility factors, and the marginal benefit of the more complex hybrid architecture.

Table 4. Classification Accuracy Across Models

Model / Input Block	Accuracy (%)
DNN – Male-only	80.2
DNN – Female-only	84.6
DNN – Combined	87.3
Hybrid CNN + LSTM – Combined	88.7

3.2 Discussion

The primary finding of this study is the significant enhancement in prognostic accuracy achieved by models that incorporate the combined fertility parameters of both male and female partners. The Hybrid CNN + LSTM – Combined model yielded the highest predictive performance (Accuracy=0.887, AUROC=0.918, Brier Score=0.134), confirming that the combined approach is necessary to maximize prognostic accuracy.

The results clearly establish a hierarchy of predictive power, directly correlating with the completeness of the input data. The standard DNN – Combined model (Accuracy=0.873, AUROC=0.912) significantly outperformed the single-factor models, demonstrating that the integration of male fertility parameters is necessary to resolve predictive uncertainty. Consistent with existing reproductive biology literature, the DNN – Female-only model (Accuracy=0.846, AUROC=0.884) substantially outperformed the DNN – Male-only model (Accuracy=0.802, AUROC=0.861), showing a $\approx4.4\ \text{pp}$ advantage in accuracy and a $\approx2.3\ \text{pp}$ advantage in AUROC. This aligns with evidence showing that female factors, particularly Age, Endometrial thickness, and AMH levels (Table 2), which relate to oocyte quality and uterine receptivity, are the central determinants of implantation success (Esteves et al., 2021).

However, the necessity of the combined input is confirmed by the performance metrics and feature importance. The male factors contributed significantly (approximately $\text{31}\%$ of explained variance) in the combined models. The emergence of Sperm morphology, Progressive motility, and the DNA-fragmentation index as critical predictors underscores the clinical relevance of comprehensive male evaluation. These findings are consistent with prior work by

Barragán et al. (2018), which linked sperm chromatin integrity to subsequent embryo development, reinforcing those male parameters are indispensable, particularly in cases of idiopathic infertility. The superior performance of the Hybrid CNN + LSTM architecture over the simpler DNN – Combined model (an incremental gain of AUROC =0.006 suggests that the hybrid approach is slightly better at capturing subtle, non-linear dependencies in the combined feature set. The resulting low Brier Score confirms the superior calibration and reliability of the Hybrid model's probability predictions.

Table 4 succinctly summarizes the core finding regarding model accuracy. The observed trend, where accuracy systematically increases from single-factor (male-only, female-only) to combined-factor models, and then slightly improves with a more complex architecture (Hybrid CNN + LSTM), provides strong evidence for several key points. Firstly, the performance gap between the DNN – Male-only (80.2\%) and DNN – Female-only (84.6%) models highlights the dominant role of female fertility factors in IVF success prediction, aligning with existing biological understanding that places significant emphasis on oocyte quality and maternal uterine environment. However, the subsequent, more substantial increase in accuracy when combining male and female data (from84.6% for DNN – Female-only to 87.3% for DNN – Combined) underscores the indispensable contribution of male factors. This jump of 27% in accuracy demonstrates that male parameters provide unique and critical information that significantly enhances the overall predictive power, even if female factors individually appear to be stronger predictors. Finally, the marginal but notable improvement observed with the Hybrid CNN + LSTM – Combined model (88.7%) over the simpler DNN – Combined model suggests that advanced deep learning architectures can capture more intricate, non-linear relationships within the integrated dataset, leading to subtle but valuable gains in prediction accuracy. This systematic increase in accuracy across the models validates the multi-factorial nature of IVF success and reinforces the technical advantages of comprehensive data integration and sophisticated model design.

3.2.1 Implications for Clinical Practice

The deployment of deep learning (DL) models in assisted reproductive technology (ART), especially in in vitro fertilization (IVF), presents a transformative opportunity for clinical workflows. Our findings demonstrate that AI systems can effectively integrate heterogeneous fertility data to predict IVF outcomes with high accuracy and reliability. This supports the integration of such models as Clinical Decision Support Tools (CDSTs), allowing embryologists and fertility specialists to stratify patient risk for implantation failure or cycle cancellation. These tools can optimize treatment protocols by tailoring stimulation, insemination, and embryo transfer strategies based on the individual's combined fertility profile. Furthermore, they can reduce subjectivity in embryo selection and partner fertility assessment, supplementing expert judgment with consistent, data-driven insights. This technology can also enhance counseling by providing probabilistic outcome forecasts, enabling more informed consent and better emotional preparedness for patients undergoing IVF.

3.2.2 Generalizability and Dataset Bias

While our results are highly promising for the development of prediction models, generalizability is constrained by potential dataset biases. The geographic and demographic concentration of the training data, primarily originating from a small number of fertility centers in Europe and North America, limits the direct applicability of the model to other populations, including African and Asian cohorts. Furthermore, the underrepresentation of certain subfertility phenotypes such as polycystic ovary syndrome (PCOS), varicocele, or unexplained infertility may skew predictions when the model is applied to a more diverse clinical population. The absence of ethnic and socioeconomic diversity in training data also poses a risk of introducing algorithmic bias, potentially affecting fairness and accuracy across different subpopulations. Efforts to develop

global, federated IVF datasets and validate models across multi-ethnic cohorts are necessary steps for responsible and ethical deployment.

3.2.3 Limitations and Future Work

Despite achieving state-of-the-art performance, several limitations must be acknowledged. The dataset size (n=1,112 IVF cycles), although statistically adequate, may not fully capture all clinical variations; thus, a larger, multicenter dataset would enhance model robustness. Our current model predicted implantation success but lacked longitudinal outcome data, failing to track pregnancy progression or live birth rates. Future models should incorporate these critical longitudinal outcomes, including early miscarriage and neonatal health. Regarding the model itself, while SHAP values provided some transparency, the black-box nature of deep learning still poses challenges, necessitating future work to explore explainable AI (XAI) techniques such as attention mechanisms or counterfactual reasoning. Finally, translating these AI models into real-time clinical tools will require addressing challenges in regulatory validation, user interface design, and seamless integration into Electronic Health Record (EHR) systems. Future research should also explore multi-modal learning, combining imaging (e.g., embryo morphology), genomics (e.g., PGT-A), and clinical data for more holistic fertility prediction.

4.0 Conclusion

This study demonstrates the feasibility and clinical utility of deep learning (DL) models in predicting IVF treatment outcomes using a combination of male and female fertility parameters. Our findings showed that while female factors such as age, endometrial thickness, and AMH levels were dominant predictors, male factors—including sperm morphology and DNA fragmentation—also made substantial contributions to predictive performance. Importantly, models that incorporated both partners' data significantly outperformed single-gender input pipelines, reinforcing the need for a couple-focused approach in fertility assessment and treatment planning. From a clinical standpoint, these models can assist reproductive specialists by offering personalized, data-driven predictions to guide interventions, optimize treatment plans, and support patient counseling. They also offer potential for reducing subjective biases in embryo selection and partner evaluation, ensuring consistency in clinical decisions.

5.0 References

Esteva, A., Robicquet, A., Ramsundar, B., Kuleshov, V., DePristo, M., Chou, K., … Dean, J. (2019). A guide to deep learning in healthcare. *Nature Medicine*, *25*(1), 24–29. https://doi.org/10.1038/s41591-018-0316-z

European Society of Human Reproduction and Embryology (ESHRE). (2020). ESHRE guidelines for the evaluation of infertility. *Human Reproduction Open*, *2020*(1), hoaa005. https://doi.org/10.1093/hropen/hoaa005

Jindal, S. K., & Gupta, A. (2019). Artificial intelligence applications in reproductive medicine. *Journal of Human Reproductive Sciences*, *12*(3), 177–182. https://doi.org/10.4103/jhrs.JHRS_119_18

Kawwass, J. F., Crawford, S., Kissin, D. M., Jamieson, D. J., & Session, D. R. (2020). Trends and outcomes for assisted reproductive technology cycles in the United States. *Fertility and Sterility*, *113*(3), 521–529. https://doi.org/10.1016/j.fertnstert.2019.10.022.

States. *Fertility and Sterility*, *113*(3), 521–529. https://doi.org/10.1016/j.fertnstert.2019.10.021

Nwobodo, E. I., & Isah, A. Y. (2017). Semen quality in infertile Nigerian men: A prospective study. *Nigerian Journal of Clinical Practice*, *20*(4), 475–479. https://doi.org/10.4103/1119-3077.187319

Obiechina, N. J. A., & Okafor, C. N. (2018). Artificial intelligence in reproductive health: Implications for Nigeria. *International Journal of Medicine and Health Development*, *23*(2), 98–104.

Oladokun, A., Sule-Odu, A., & Adeniji, A. (2020). Patterns of male factor infertility in a Nigerian tertiary hospital. *African Journal of Reproductive Health, 24*(2), 81–88. https://doi.org/10.29063/ajrh2020/v24i2.9.

Declarations

Ethics and Consent to Participate

Not applicable.

Consent to Publish

Not applicable

Availability of data and materials

The datasets used or analyzed during the current study are available from the corresponding author upon reasonable request.

Funding

The authors declared no external source of funding

Competing Interests

The authors have no relevant financial or non-financial interests to disclose.

Authors' Contributions

E.A.E. conceptualized the study, designed the deep learning framework, and supervised model implementation. O.C.B.O. handled data collection, preprocessing, and statistical analysis. P.E. developed and optimized the hybrid CNN-LSTM architecture, validated results, and contributed significantly to manuscript writing and result interpretation.