Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

MIDLPred: Ensemble Learning for Phenotype-Oriented Classification of Male Infertility Candidate Proteins from Primary Amino Acid Sequences

Abdelhafedh BEN DAHMANE

abdelhafedh.ben-dahmane@doc.umc.edu.dz

University Frères Mentouri Constantine 1

Mustapha BENSAADA

University Frères Mentouri Constantine 1

Hamza CHEHILI

University Frères Mentouri Constantine 1

Salah eddine ALIOUANE

University Frères Mentouri Constantine 1

Mohamed Abdelhafid HAMIDECHI

University Frères Mentouri Constantine 1

Research Article

Keywords: Bioinformatics, Candidate protein prioritization, Deep learning, Ensemble learning, Male infertility, MIDLPred, Natural language processing, Precision medicine, Protein sequence analysis

Posted Date: August 5th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10059812/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Title: MIDLPred: Ensemble Learning for Phenotype-Oriented Classification of Male Infertility Candidate Proteins from Primary Amino Acid Sequences

Abdelhafedh BEN DAHMANE[1,2,*], Mustapha BENSAADA[1,3], Hamza CHEHILI[1,4], Salah eddine ALIOUANE[1,2], Mohamed Abdelhafid HAMIDECHI[1,3]

[1] Department of Applied Biology, Constantine 1 University, Constantine 25000, Algeria.

[2] Microbiological Engineering and Applications Laboratory, Constantine 1 University, Constantine 25000, Algeria

[3] Mycology, biotechnology and microbial activity laboratory, Bioinformatics and Artificial Intelligence Research Team (B2IA), Constantine 1 University, Constantine 25000, Algeria.

[4] LIRE Laboratory, Constantine 2 University, Constantine 25000, Algeria

* Correspondence: abdelhafedh.ben-dahmane@doc.umc.edu.dz

ABSTRACT

Background: Male infertility is a complex and multifactorial condition resulting from the interplay of numerous genetic, molecular, physiological, and environmental factors. Advances in reproductive genetics have identified a growing number of genes and proteins involved in spermatogenesis, sperm motility, sperm morphology, chromatin organization, and testicular function. However, linking candidate proteins to clinically relevant infertility phenotypes remains a major challenge in molecular andrology and reproductive medicine.

Objectives: The objective of this study was to develop a computational framework capable of identifying phenotype-associated molecular signatures directly from protein sequences and assisting in the prioritization of candidate proteins involved in male infertility.

Materials and Methods: A dedicated dataset was generated through a bioinformatics workflow integrating information from the Male Infertility Knowledgebase (MIK), UniProt, and complementary biological resources. The initial dataset comprised 31,774 protein entries associated with male infertility. Following sequence quality assessment, duplicate removal, annotation harmonization, and filtering procedures, 10,678 curated protein sequences were retained. To facilitate biologically meaningful analysis, proteins were grouped into four phenotype-oriented categories: MOTILITY, MORPH, PHYSIO, and OTHER. A one-dimensional convolutional neural network (1D-CNN) combined with an ensemble learning strategy involving ten independently trained models was developed to learn phenotype-related sequence patterns. Model performance was assessed using stratified cross-validation, confusion matrix analysis, receiver operating characteristic (ROC) curves, and standard classification metrics.

Results: MIDLPred achieved an overall classification accuracy of 96% and a weighted F1-score of 96% across the four-phenotype categories. Cross-validation analyses demonstrated stable and reproducible performance, while confusion matrix and ROC analyses confirmed the precision of the proposed framework. The model successfully identified sequence-derived patterns associated with distinct infertility phenotypes, suggesting that biologically relevant information is embedded within primary protein sequences.

Discussion and conclusion: The results demonstrate the potential of deep learning approaches for exploring the molecular landscape of male infertility through protein sequence analysis. Rather than serving as a standalone diagnostic tool, MIDLPred provides a bioinformatics framework that may assist clinicians, reproductive biologists, geneticists, and andrologists in prioritizing candidate proteins and generating biologically relevant hypotheses for further investigation. These findings highlight the value of integrating artificial intelligence with reproductive genetics and molecular andrology and support future developments incorporating explainable artificial intelligence, protein language models, and multi-omics data to improve biological interpretation and clinical applicability.

MIDLPred is publicly available at https://midlpred-amob.onrender.com/

Keywords: Bioinformatics, Candidate protein prioritization, Deep learning, Ensemble learning, Male infertility, MIDLPred, Natural language processing, Precision medicine, Protein sequence analysis.

INTRODUCTION

Male infertility represents a critical global public health concern, impacting approximately 7% of the male population worldwide [1]. Historically, the World Health Organization (WHO) has established criteria for assessing male fertility, primarily based on semen analysis [2] . Despite their widespread use, these criteria exhibit notable shortcomings and often fail to deliver accurate prognostic insights [3]. These limitations are further underscored by recent expert recommendations, such as those from the French BLEFCO group, which advocate for a significant simplification of sperm morphology assessment and challenge its systematic prognostic value for assisted reproductive techniques (ART)[4].

In the context of precision medicine, the demand for more sophisticated diagnostic strategies has intensified, particularly given the intricate nature of omics data.

This is exemplified by the application of whole-exome sequencing, which not only improves the genetic diagnosis of severe conditions like non-obstructive azoospermia but can also predict the success of surgical sperm retrieval, thereby guiding clinical decisions and avoiding unnecessary interventions [5]. To meet this challenge, there is a growing need for computational tools that not only increase processing efficiency but also enhance predictive performance.

This study explores the development of a classification system for male infertility leveraging two powerful branches of artificial intelligence: Natural Language Processing (NLP) and Deep Learning (DL). Among these, deep learning stands out due to its demonstrated success across diverse biomedical domains [6]

Advancements in proteomics and AI have opened new avenues for improving the diagnosis and understanding of male infertility [7,8].

Our approach applies a deep learning framework to protein expression data, complemented by genomic insights, to identify key determinants of reproductive dysfunction. Recent advances further demonstrate that deep learning architectures leveraging protein language model embeddings can effectively predict structural and functional characteristics directly from protein sequences, including intrinsic and soft disorder regions[9]

Through analysis of validated datasets, we found that genes associated with male infertility are distributed across multiple chromosomal regions, some of which appear to be recurrently enriched. Furthermore, many of these regions are implicated in multiple reproductive disorders, reflecting a pleiotropic genetic architecture that complicates the interpretation of infertility-related phenotypes [10].

These findings are illustrated in the Heatmap (Figure 1), which maps the chromosomal localization and shared associations of relevant genes [10].

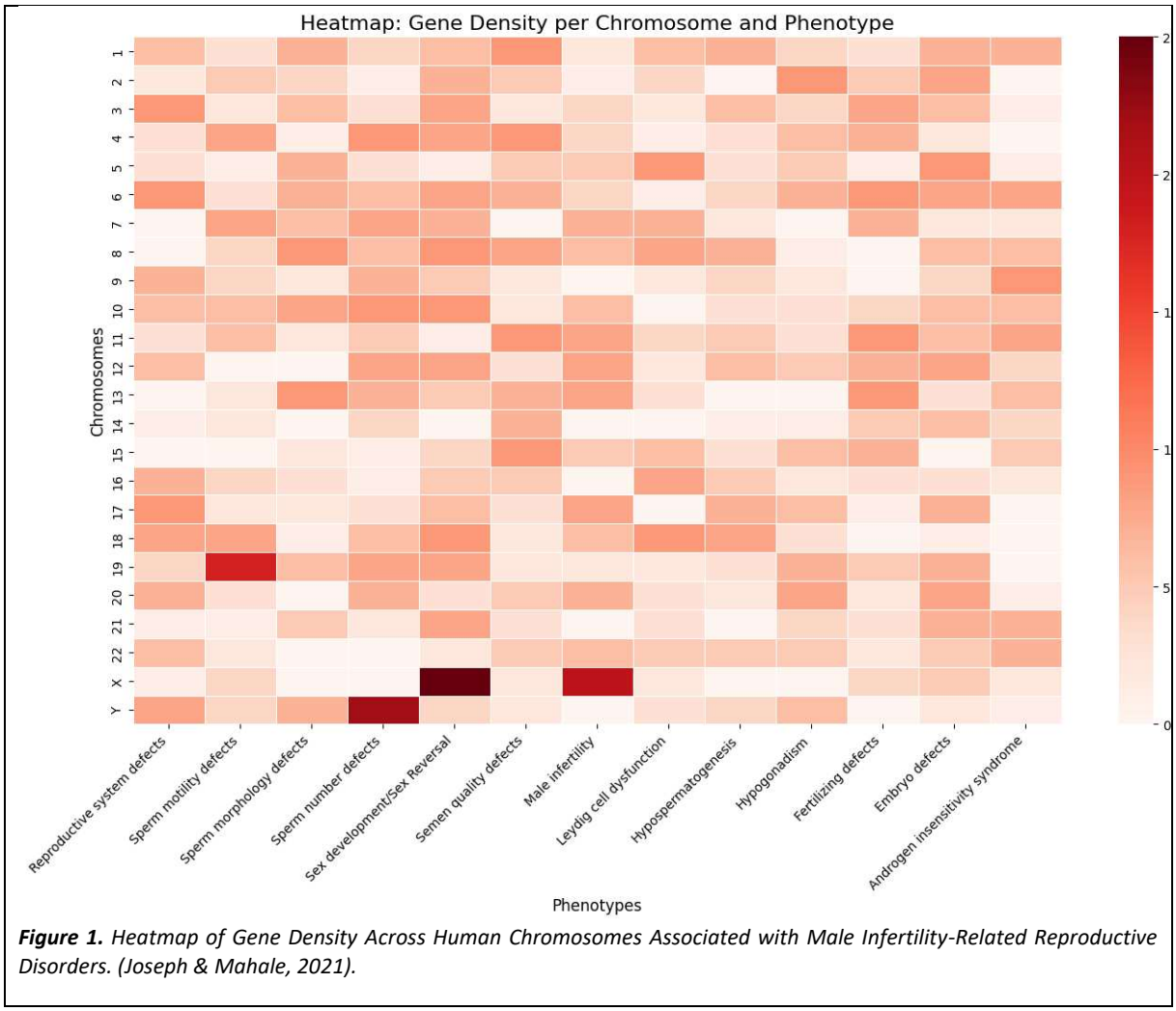


Figure 1. *Heatmap of Gene Density Across Human Chromosomes Associated with Male Infertility-Related Reproductive Disorders. (Joseph & Mahale, 2021).*

By analyzing protein signatures in conjunction with these genomic patterns, our method provides a more refined and individualized alternative to existing WHO guidelines [11].

Deep learning enables the interpretation of complex biological relationships, the discovery of novel biomarkers, and improved patient stratification. Ultimately, this lays the groundwork for more precise diagnostic processes and the development of targeted therapeutic strategies [8,12].

In this article, we detail the design and implementation of our AI-driven model, its application to large-scale proteomic and genomic datasets, and its potential implications for advancing the diagnosis and treatment of male infertility [13].

MATERIAL AND METHODS

Study subjects, and Research Hypotheses

Male infertility is rarely caused by a single molecular defect and is instead characterized by a complex interplay of genetic, molecular, developmental, and environmental factors. Nevertheless, many genes implicated in male infertility encode proteins exhibiting strong or preferential expression within the testis and reproductive tissues. These proteins participate in specific biological processes including spermatogenesis, sperm motility, sperm morphology, chromatin organization, hormonal regulation, and testicular development.

Recent advances in protein sequence analysis and deep learning suggest that biologically meaningful information can be extracted directly from amino acid sequences without requiring explicit structural annotations. We therefore hypothesized that proteins associated with distinct infertility phenotypes may contain phenotype-specific sequence signatures that can be learned by modern deep learning architectures.

The present work was designed to evaluate three hypotheses:

- H1: Protein sequences associated with male infertility contain phenotype-specific molecular signatures that can be captured by deep learning models.
- H2: Ensemble learning improves prediction precision and generalization compared with individual neural network models.
- H3: Sequence-based phenotype prediction can assist clinicians, reproductive biologists, and andrologists in prioritizing candidate proteins potentially involved in male infertility mechanisms.

Rather than functioning as a standalone diagnostic system, MIDLPred aims to serve as a computational decision-support framework capable of guiding molecular investigations and facilitating the interpretation of candidate proteins identified through genomic and proteomic studies.

Process of the global approach

The proposed approach is organized into two primary phases: Data Preprocessing and Model Training. The overall methodology, illustrated in Figure 2, is further broken down into three interconnected phases: Data Preprocessing, Learning, and Prediction.

Data Preprocessing Phase

The initial phase, highlighted in blue in Figure 2, begins with data collection. This is followed by comprehensive data preprocessing to ensure the dataset is clean, structured, and optimized for the subsequent learning stage.

This step is critical, as the quality of data preprocessing directly impacts the model's accuracy and reliability, a principle underscored in automated analysis frameworks for biological data[14,15].The

figure 2 provides a clear overview of the proposed approach, breaking it down into three key phases: Data Preprocessing, Learning, and Prediction.

Learning Phase

The core of the approach, outlined in red in Figure 2, is the Learning phase. This phase starts with the selection of the most suitable model architecture. The selection is informed by the demonstrated strengths of Deep Learning (DL) models, particularly Convolutional Neural Networks (CNNs), in analyzing complex morphological and histological data[15,16].

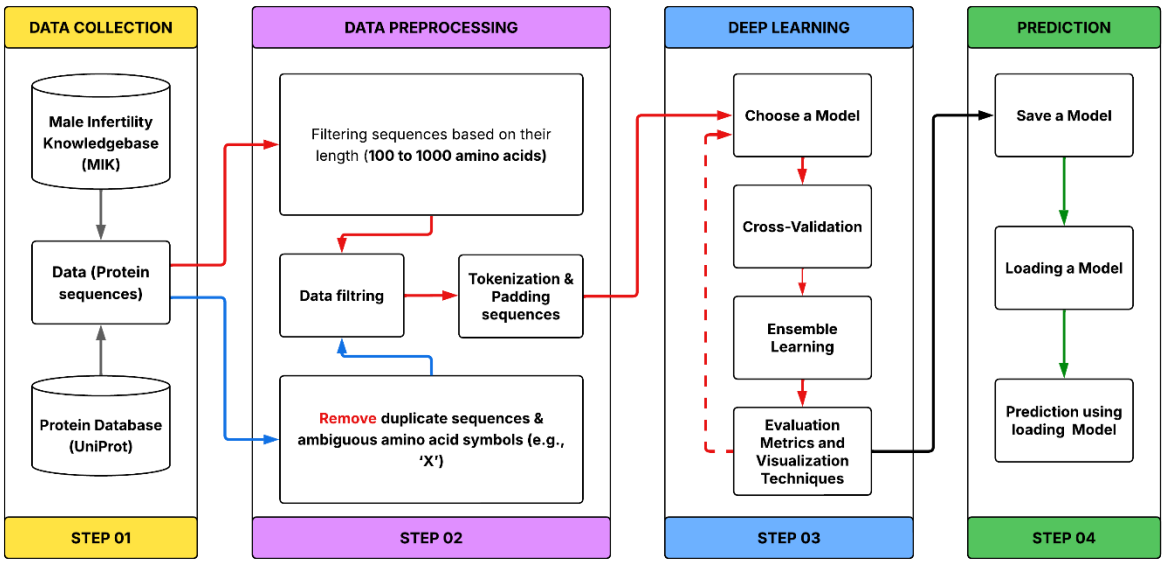


Figure 2. *Process of the global approach*

Following model selection, the process moves to the learning and assessment stage, where the model is trained and its performance is rigorously evaluated. If the initial performance is not satisfactory, hyper-parameters are fine-tuned to optimize results. This stage enables researchers to develop, refine, and customize AI models efficiently according to their specific requirements [16]. Once the model achieves a satisfactory level of accuracy, it is saved for future deployment.

Prediction Phase

The final phase, outlined in black in Figure 2, is Prediction. Here, the saved model is loaded and utilized to make predictions on new, unseen data. This demonstrates the practical application and utility of the trained model, fulfilling the ultimate goal of enhancing performance in analytical tasks, as explored in the context of applying deep learning to practical medical and biological challenges[16,17].

By adhering to this structured, three phase approach: Data Preprocessing, Learning, and Prediction, the model is systematically developed, fine-tuned, and deployed, ensuring robust and efficient predictive performance.

DATA COLLECTION

Unlike studies relying exclusively on existing benchmark datasets, we constructed a dedicated dataset for the development of the MIDLPred model through a bioinformatics workflow integrating information from the Male Infertility Knowledgebase (MIK) [10] URL: http://mik.bicnirrh.res.in/ (PMID: 34363073), UniProt, and complementary biological resources (Figure 2). Gene, protein, and male infertility phenotype information were first extracted from MIK. Custom Python scripts, database queries, and API-based retrieval procedures were then employed to synchronize these records with the UniProt database using UniProt accession numbers, enabling the retrieval of manually reviewed protein sequences and their associated annotations.

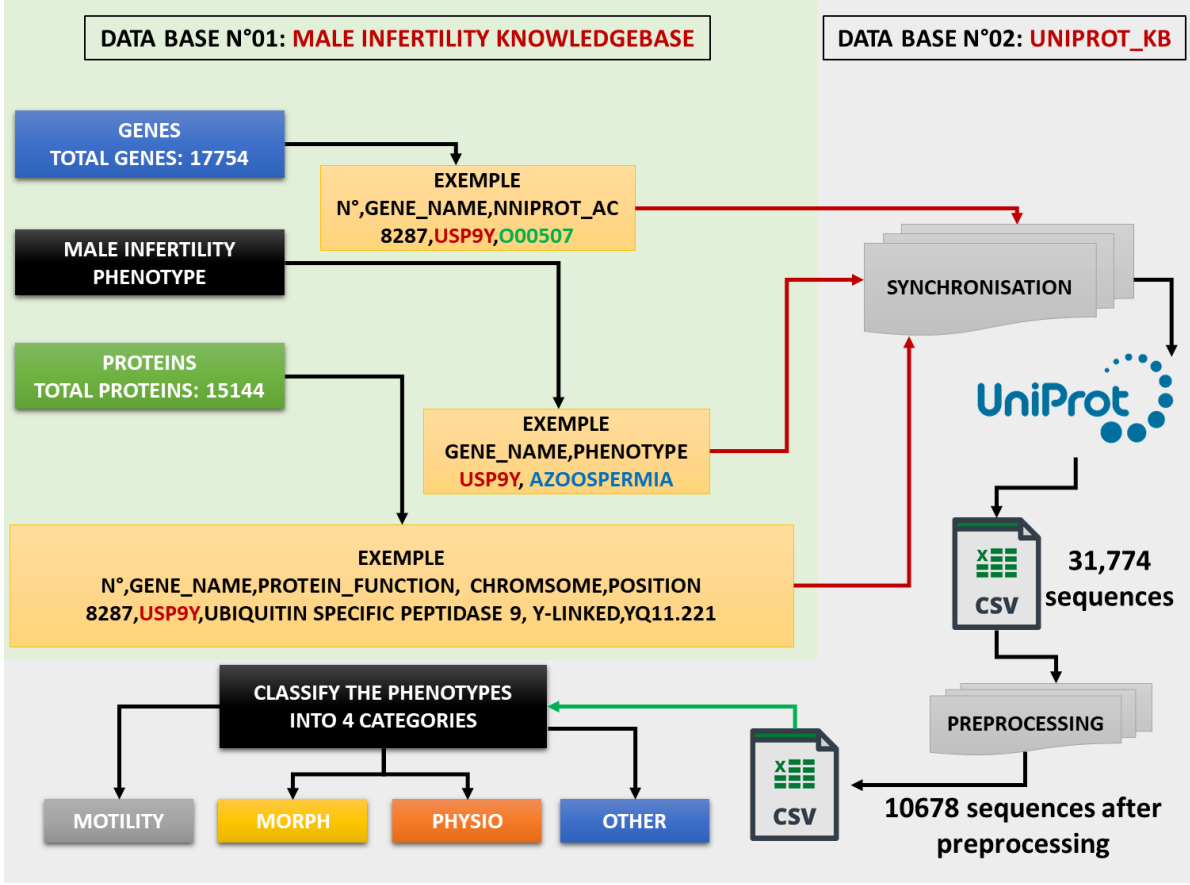


Figure 3. *Workflow of the MIDLPred dataset construction from the Male Infertility Knowledgebase (MIK) and UniProt.*

Protein sequences were retrieved from UniProtKB using the original UniProt accession identifiers associated with each MIK entry. The initial dataset comprised 31,774 protein entries related to male infertility, exported in CSV format. However, the raw collection contained redundant protein entries, synonymous records, isoforms, sequence variants arising from amino acid substitutions, and multiple annotations describing closely related infertility phenotypes. Therefore, several preprocessing, harmonization, and quality-control procedures were applied to remove inconsistencies and standardize the dataset before model development. This workflow resulted in a large-scale, curated

protein sequence dataset specifically designed for male infertility phenotype prediction and served as the foundation for training and evaluating the MIDLPred model.

Data Preprocessing

The initial dataset used in this study was derived from a curated dataset comprising protein sequences associated with male infertility factors. The raw dataset included multiple redundant or incomplete entries, necessitating a comprehensive cleaning and refinement process to ensure data quality and consistency for deep learning applications.

We began by filtering sequences based on their length, retaining only those ranging from 100 to 1000 amino acids. This threshold was selected to capture biologically relevant fragments while excluding extremely short or abnormally long sequences that might distort the learning process. After applying this filter, we identified and removed 21,096 duplicate sequences, preserving only the first instance of each unique sequence. This resulted in a refined dataset containing 10,678 unique entries.

Next, all records with missing values whether in the sequence or classification field were excluded. This yielded a final count of 10,678 protein samples with complete annotations, ensuring a clean input for downstream processing.

To facilitate biological interpretation, the dataset was curated and grouped into four clinically and biologically meaningful categories related to male infertility: MOTILITY, MORPH, PHYSIO, and OTHER (Figure 4).

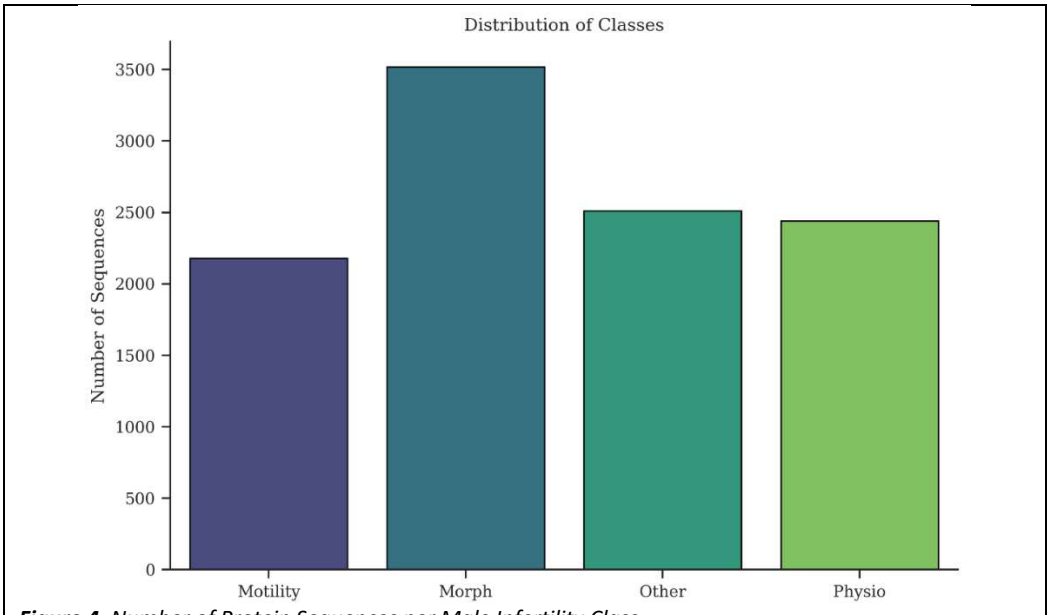


Figure 4. *Number of Protein Sequences per Male Infertility Class*

The original Male Infertility Knowledgebase contains a large number of infertility-associated phenotypes with highly heterogeneous representation. Several phenotypes were represented by only a limited number of proteins, making direct phenotype level classification statistically challenging and potentially biased. Furthermore, many infertility phenotypes share overlapping molecular mechanisms

and biological pathways. To address these limitations while preserving biological interpretability, related phenotypes were grouped into four higher-level phenotype categories.

Specifically, 3,520 protein sequences were assigned to the MORPH category, which includes morphological abnormalities such as teratozoospermia, globozoospermia, macrozoospermia, and other disorders affecting sperm structure and morphology.

The MOTILITY category comprised 2,192 protein sequences associated with sperm movement defects, including asthenozoospermia, oligoasthenozoospermia, and other conditions affecting sperm motility and flagellar function.

The PHYSIO category included 2,447 protein sequences related to testicular physiology and spermatogenic dysfunction, including azoospermia, spermatogenic failure, maturation arrest, Sertoli-cell-only syndrome, and other disorders affecting testicular function and sperm production.

Finally, 2,519 protein sequences were grouped within the OTHER category, which encompasses heterogeneous infertility-associated conditions not directly linked to sperm morphology or motility, including cryptorchidism, hypospadias, endocrine disorders, developmental abnormalities, and unexplained infertility.

This grouping strategy was designed to reduce class sparsity while allowing rare infertility phenotypes to contribute to model training through biologically coherent phenotype groups. Therefore, the objective of MIDLPred is not to predict a specific clinical diagnosis but rather to identify the infertility phenotype category most strongly associated with a candidate protein sequence.

The final curated dataset contained 10,678 unique protein sequences distributed across the four-phenotype categories. This phenotype-oriented classification framework enables proteins involved in diverse infertility manifestations to participate in a unified prediction task while maintaining biological relevance and improving statistical precision for deep learning applications (Figure 5).

These four categories were chosen based on both clinical relevance and sufficient representation within the dataset, ensuring balanced model training and reliable evaluation. The resulting dataset had a shape of (10,678, 4), representing 10,678 labeled protein sequences suitable for supervised learning.

For modeling, the dataset was split into training and test subsets using stratified sampling, allocating 90% for training and 10% for testing. This stratification preserved the proportional representation of each class across both sets, ensuring balanced evaluation.

To prepare sequences for neural network input, a character-level tokenizer was applied. Each amino acid in the sequence was treated as a distinct token, and the tokenizer was fitted exclusively on the training data to prevent information leakage. Sequences were then transformed into integer sequences and padded to a fixed length of 1,000 residues using post padding to ensure uniform input dimensions.

Class labels were encoded using a LabelEncoder, mapping each of the four infertility-related classes to a unique integer.

These pre-processed sequences and labels served as the foundation for model training and evaluation.

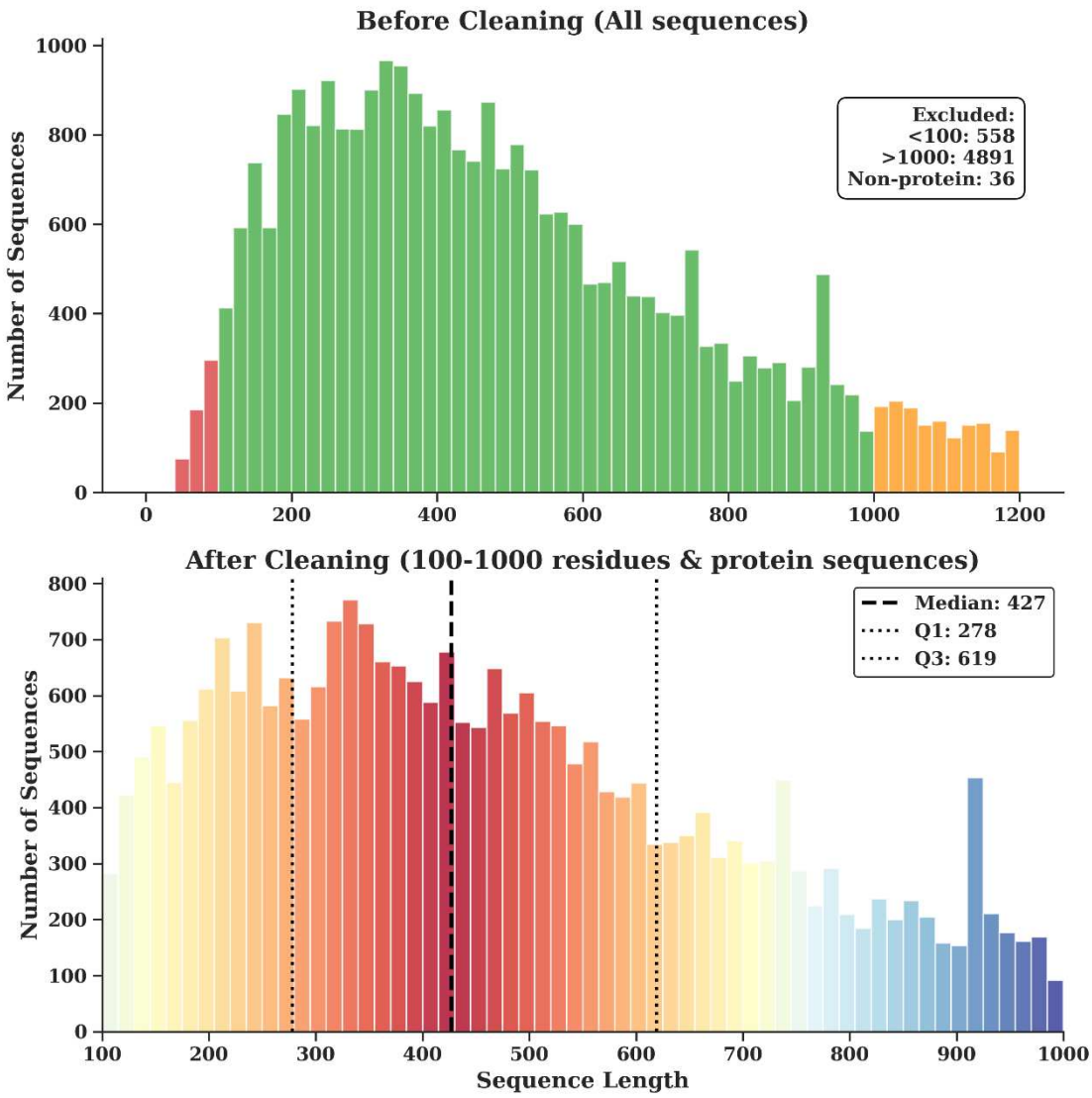


Figure 5. *Sequence Length Distribution before & after cleaning*

MODEL ARCHITECTURE

To classify protein sequences according to their associated male infertility factors, we developed a deep learning architecture based on a 1D Convolutional Neural Network (1D-CNN). This architecture was specifically chosen for its ability to capture local dependencies and patterns in biological sequences, making it well-suited for analyzing amino acid chains of variable structure (figure 6).

Each input sequence composed of up to 1,000 amino acids was first passed through an embedding layer, which transformed the integer-encoded tokens into a dense 128-dimensional representation.

This layer served to map each amino acid to a continuous vector space, capturing semantic and contextual relationships among residues.

The embedding layer effectively acts as a trainable look-up table, allowing the model to learn task-specific amino acid representations during training.

Following the embedding, the architecture incorporated three stacked convolutional blocks, each consisting of a 1D convolutional layer followed by a max-pooling operation:

- The first convolutional layer applied 128 filters with a kernel size of 3 and ReLU activation to capture low-level motifs.
- The second and third convolutional layers, also using 128 filters, allowed the network to learn increasingly abstract sequence features.
- Each convolutional operation was followed by a MaxPooling1D layer with a pool size of 2, which reduced the spatial dimensions of the feature maps, helping the model generalize by focusing on dominant features and reducing overfitting.

The output of the final convolutional block was flattened and passed into a series of fully connected (dense) layers:

- The first dense layer comprised 256 units, activated by ReLU, followed by a Dropout layer (rate = 0.5) to reduce overfitting by randomly deactivating neurons during training.
- Subsequent dense layers with 128, 64, and 64 units respectively were also activated using ReLU and interleaved with additional Dropout layers (each with a 0.5 dropout rate). These layers enabled the network to learn complex nonlinear relationships between extracted sequence features and the infertility classification task.
- The final output layer was a Dense layer with 4 neurons, each corresponding to one of the male infertility categories (MOTILITY, MORPH, PHYSIO, OTHER). A Softmax activation function was applied to this layer to produce a normalized probability distribution over the four classes, enabling multi-class classification.

The model was compiled using the Adamax optimizer, a variant of the Adam algorithm known for its stability in sparse data environments. A learning rate of 0.001 was selected based on empirical tuning. The loss function used was sparse categorical crossentropy, which is appropriate when target labels are encoded as integers rather than one-hot vectors. The network was trained to minimize this loss while optimizing accuracy as the primary performance metric.

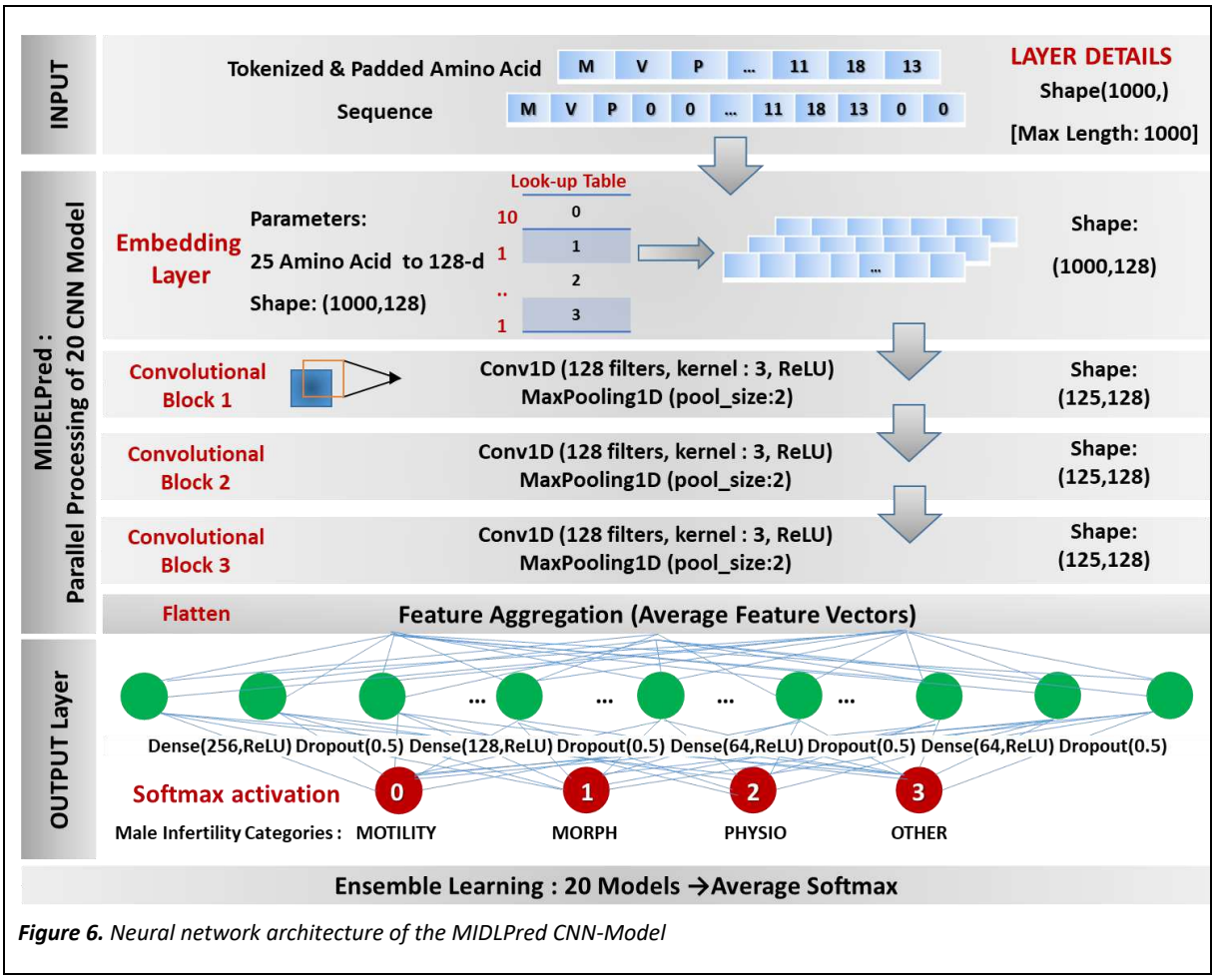


Figure 6. *Neural network architecture of the MIDLPred CNN-Model*

The architecture was implemented using the Keras deep learning library with a TensorFlow backend, ensuring modular design and scalability for future extensions or integration with additional biological features.

This carefully designed 1D-CNN framework provided a robust, biologically informed foundation for learning from protein sequence data and predicting the underlying infertility factors with high interpretability and generalization potential (Table 1).

Tableau 1. The hyper-parameters used to train the model

Layer (type)	**Output Shape**	**Param#**
embedding (Embedding)	(None, 800, 24)	504
conv1d (Conv1D)	(None, 798, 32)	2336
max_pooling1d	(None, 399, 32)	0
dense (Dense)	(None, 399, 32)	1056
dropout (Dropout)	(None, 399, 32)	0
flatten (Flatten)	(None, 12768)	0
dense_1 (Dense)	(None, 4)	51076

Total params:	54972	(214.73 KB)
Trainable params:	54972	(214.73 KB)
Non-trainable params:	0	(0.00 Byte)
None		

Cross-Validation

To assess the generalization capability of our model and ensure that performance estimates were robust to sampling variability, we employed a five-fold stratified cross-validation strategy on the training dataset. This method partitions the dataset into five mutually exclusive folds while preserving the relative class distribution across each fold, which is critical in multi-class classification tasks with potential class imbalance.

In each cross-validation iteration, the model was trained on four of the five folds (80%) and validated on the remaining fold (20%). This process was repeated five times, with a different fold held out for validation during each iteration. As a result, every sample in the training set was used exactly once for validation, ensuring an unbiased estimation of the model's performance on unseen data.

Each fold was trained using the same model architecture described previously, with consistent hyperparameters and regularization settings. Training was conducted over 50 epochs, with a batch size of 64, and monitored using three callbacks:

- EarlyStopping, configured with a patience of 20 epochs, halted training if the validation loss did not improve, preventing overfitting.
- ReduceLROnPlateau, which adaptively reduced the learning rate by a factor of 0.9 when validation loss plateaued, with a minimum learning rate of 1e-6.
- ModelCheckpoint, which saved the model weights achieving the lowest validation loss for each fold.

After each training cycle, the model was evaluated on the validation set using accuracy, precision, recall, and F1-score, computed using weighted averages to account for class imbalance. The confusion matrix and classification report were also generated for each fold to gain detailed insight into the performance across the four infertility categories (MOTILITY, MORPH, PHYSIO, OTHER).

This cross-validation process not only mitigated the risk of overfitting to a specific partition but also provided a comprehensive picture of the model's behavior across different data splits. The resulting distribution of scores across the five folds served as a reliable indicator of performance variability and informed the ensemble learning strategy described in the following section.

Ensemble Learning

To further enhance predictive performance and reduce variance introduced by individual model training, we implemented an ensemble learning approach composed of ten independently trained 1D-CNN models. Each model was trained on a different stratified split of the training data, introducing diversity in learning while maintaining consistent class distributions.

The ensemble was constructed as follows: the full training set (90% of the preprocessed dataset) was repeatedly split into training and validation subsets using an 80/20 ratio. Each split was used to train a

separate instance of the model architecture described previously. All models shared the same hyperparameters, optimizer (Adamax, learning rate = 0.001), and training configurations, including EarlyStopping, ReduceLROnPlateau, and ModelCheckpoint callbacks to ensure stable convergence and prevent overfitting.

For each model, predictions were generated on the held-out 10% test set, and performance metrics including accuracy, precision, recall, and F1-score were calculated individually. Confusion matrices and classification reports were also compiled to assess the strengths and weaknesses of each model with respect to the four male infertility classes.

Once all ten models were trained and evaluated, their output probabilities on the test set were aggregated using probabilistic averaging. Specifically, the predicted class probabilities from each model were summed and then averaged to produce the final ensemble prediction. This approach allowed the ensemble to leverage complementary decision boundaries learned by individual models, effectively smoothing out overfitting tendencies and enhancing precision.

The final class label for each test sample was determined by selecting the class with the highest averaged probability. The ensemble's performance was evaluated on the test set using the same suite of metrics as individual models. In addition to improved classification accuracy and F1-score, ensemble predictions also demonstrated higher consistency and generalization across all four classes.

ROC curves were generated for each individual model as well as for the ensemble classifier, highlighting improved micro-average AUC scores when predictions were combined. These ensemble-level results underscore the benefit of aggregating multiple trained models, especially when dealing with biological sequence data that may contain subtle patterns not consistently captured by single models.

The MIDLPred model was trained on a workstation equipped with an AMD Ryzen 5 3600X processor, 48 GB DDR4 RAM (3600 MHz), Samsung 970 SSD storage, and an NVIDIA RTX 3060 Ti GPU (8 GB VRAM). The dataset consisted of 10,678 protein sequences, and the final architecture comprised 54,972 trainable parameters. The trained models were subsequently saved as 20 files.h5 and integrated into the prediction platform to enable automated inference.

Evaluation Metrics and Visualization Techniques

To comprehensively evaluate the performance of both individual models and the ensemble classifier, we employed a set of standard classification metrics tailored for multi-class prediction tasks. The primary metrics included:

- Accuracy: the overall proportion of correct predictions across all classes.
- Precision: the ability of the model to correctly identify positive cases for each class, computed using a weighted average to account for class imbalance.

- Recall (Sensitivity): the model's capacity to capture all true instances of each class, also computed in a weighted fashion.
- F1-Score: the harmonic mean of precision and recall, providing a balanced measure of the model's classification performance.

Each of these metrics was calculated for every fold during cross-validation, for each individual model during ensemble training, and for the final ensemble prediction. By reporting both per-fold and aggregated metrics, we obtained a robust and interpretable assessment of the model's generalization performance and its consistency across different training partitions.

In addition to numerical performance measures, several visualization techniques were employed to provide further insight into the model's behavior:

Confusion Matrices were generated for each model and for the final ensemble prediction. These heatmaps illustrated true versus predicted classifications, highlighting class-specific performance and misclassification patterns. Each confusion matrix was annotated with the infertility categories (MOTILITY, MORPH, PHYSIO, OTHER) for ease of interpretation.

Receiver Operating Characteristic (ROC) Curves were plotted for each class individually, as well as using a micro-average strategy, which aggregates true positive and false positive rates across all classes. The corresponding Area Under the Curve (AUC) scores were calculated to quantify the discriminatory power of the models. ROC curves were produced for all 10 individual models as well as for the ensemble, revealing consistent performance improvements from model aggregation.

Training Curves were plotted to monitor model learning over time. For each model, we visualized the training and validation loss as well as the training and validation accuracy across all epochs. These plots provided insights into convergence behavior, potential overfitting, and the effectiveness of learning rate adaptation and early stopping.

Together, these evaluation techniques offered a multi-dimensional understanding of model performance, supporting both quantitative benchmarking and qualitative assessment.

The combination of strong predictive metrics and detailed visualizations confirmed the capacity of the MIDLPred framework to accurately and reliably classify protein sequences associated with male infertility, while offering transparency into its decision making process.

RESULTS

A cross-validation training approach was employed to robustly evaluate the model's effectiveness;

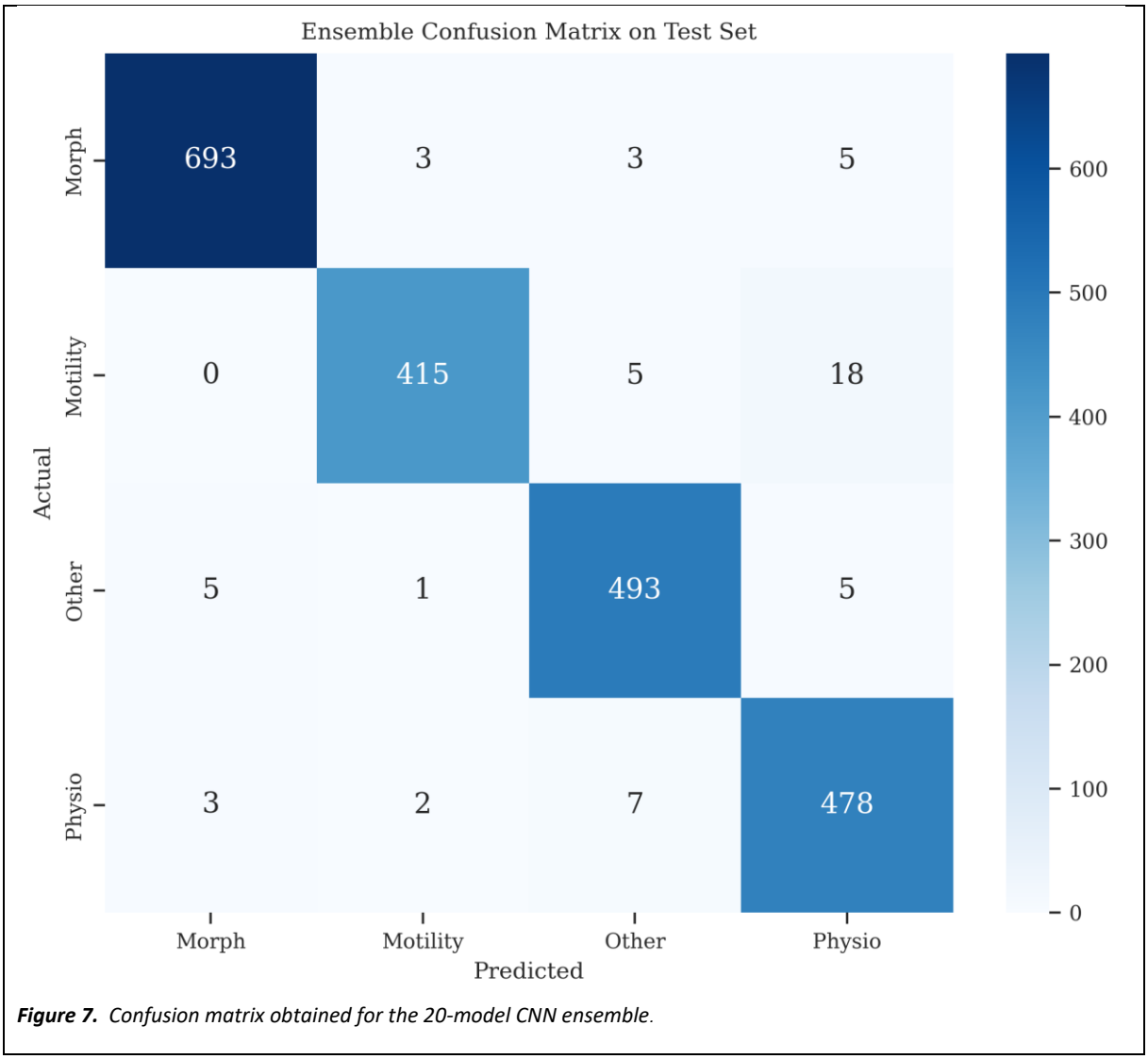


Figure 7. *Confusion matrix obtained for the 20-model CNN ensemble.*

High precision was achieved, with training precision (using cross-validation) reaching approximately 99.85% and average evaluation precision across validation folds at 99.71%. A confusion matrix (Figure 7) further assessed classification quality, visualizing the agreement between predicted and actual classifications of male infertility factors.

The strong diagonal presence in the matrix, with dark green elements, highlights the model's high accuracy in correctly classifying instances. The near-absence of off-diagonal elements signifies minimal misclassifications, demonstrating the model's reliability.

This overall structure indicates effective generalization across different classes, confirming the model's precision for classifying male infertility factors.

The learning behavior of MIDLPred is illustrated by the loss curves presented in Figure 8. The figure shows the evolution of the training and validation loss of the 20-model CNN ensemble over 50 training iterations. A rapid decrease in both loss values is observed during the early stages of training, followed by a gradual stabilization as the learning process progresses.

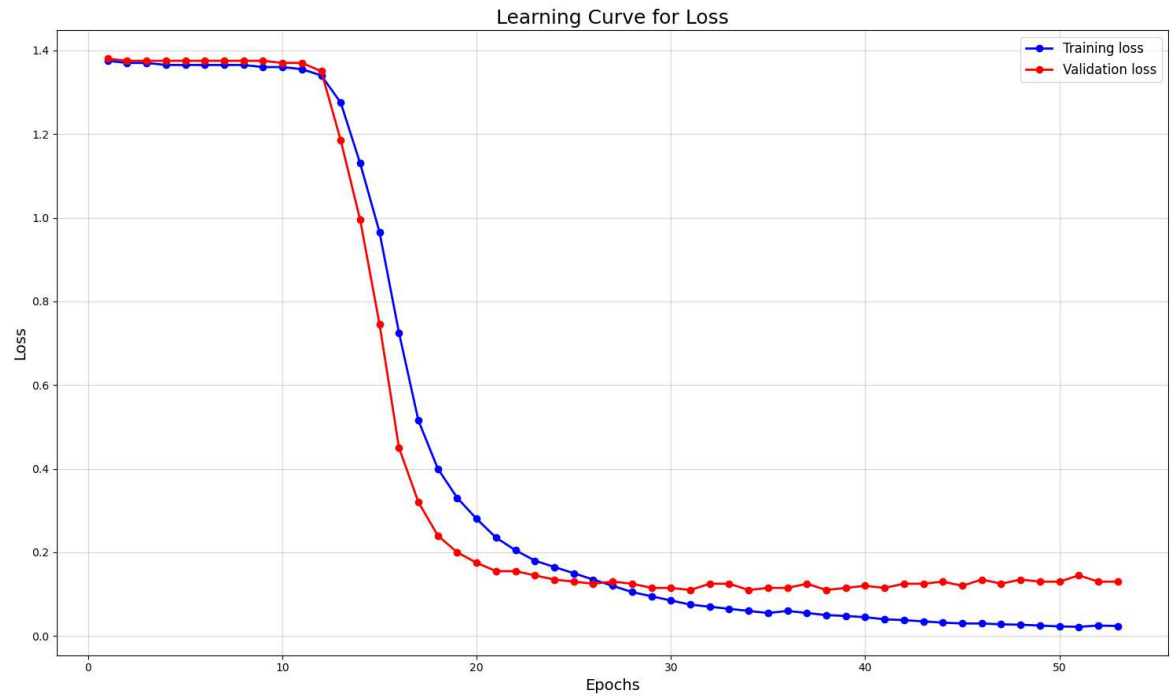


Figure 8. *Loss curve of the 20-model CNN ensemble over 50 training iterations*

The close agreement between the training and validation loss curves indicates that the ensemble model successfully learned relevant sequence features while maintaining good generalization performance. Moreover, the absence of a substantial divergence between the two curves suggests that overfitting was effectively controlled throughout training. The low and stable loss values achieved during the final iterations further support the robustness of the proposed framework and its ability to accurately classify proteins across the four male infertility phenotype categories.

Complementing the loss analysis, Figure 9 presents the precision curves of the 20-model CNN ensemble across 50 training iterations. A rapid improvement in precision is observed during the initial stages of training, followed by a stable plateau at high precision values. The close agreement between the training and validation precision curves suggests that the ensemble model effectively captures meaningful sequence patterns while maintaining strong generalization capability. Furthermore, the stability of the validation precision throughout the training process indicates that the model does not suffer from substantial overfitting. These findings are consistent with the overall classification performance reported in this study and demonstrate the ability of MIDLPred to accurately and reliably classify proteins across the four male infertility phenotype categories.

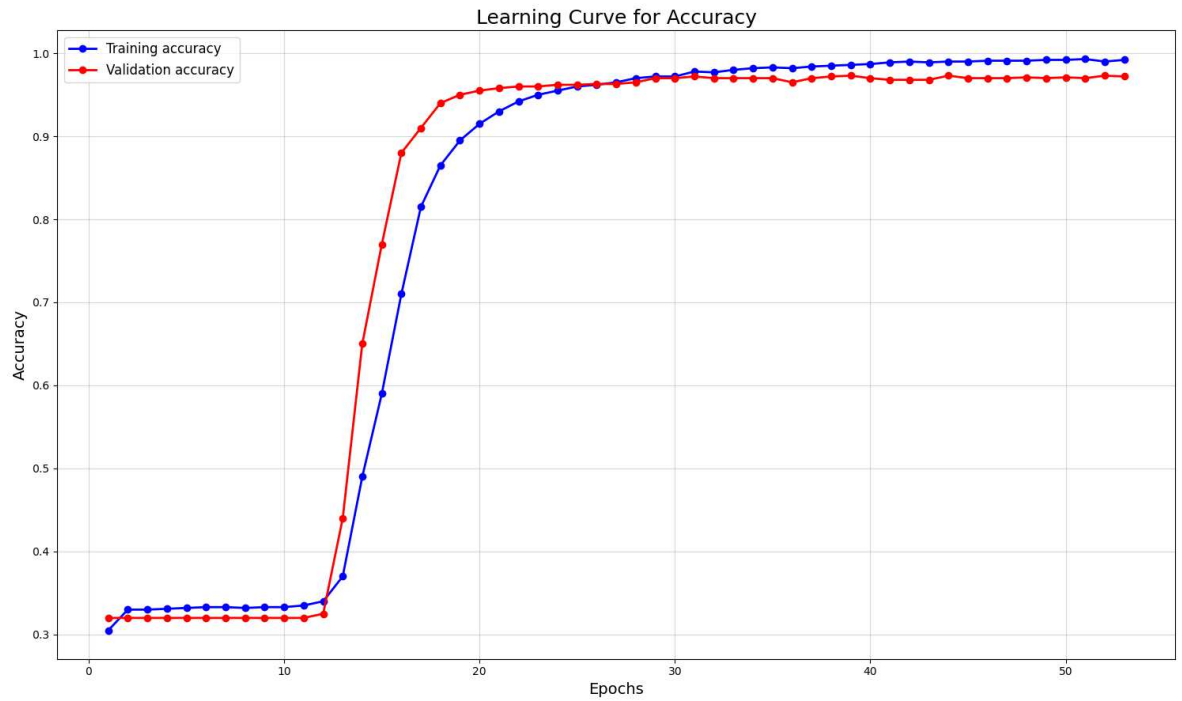


Figure 9. *Precision curve of the 20-model CNN ensemble over 50 training iterations*

A detailed breakdown of the model's performance across all classes is provided in Table 2. The classification report reveals exceptionally high and balanced metrics for each male infertility factor class Morph, Motility, Other, and Physio with precision, recall, and F1-scores all exceeding 0.93. The macro and weighted averages of 0.96 further underscore the model's consistent performance without bias towards any particular class. The high per-class scores, supported by the substantial number of sequences in the "Support" column, validate the model's effectiveness in distinguishing between the different factors, as initially suggested by the strong diagonal in the confusion matrix.

Tableau 2. Classification report showing accuracy, recall, f1 score and num-ber of sequences tested for each factor class

Class	**Precision**	**Recall**	**F1-score**	**Support**
Morph	0.98	0.99	0.98	704
Motility	0.98	0.94	0.96	438
Other	0.94	0.97	0.95	504
Physio	0.94	0.93	0.94	490
Accuracy			0.96	2136
Macro avg	0.96	0.96	0.96	2136
Weighted avg	0.96	0.96	0.96	2136

The distribution of evaluation metrics across the 20 individual models is presented in Figure 10. Accuracy, Precision, Recall, and F1-score all show remarkably narrow interquartile ranges (width ≤

0.005) and no extreme outliers, demonstrating the stability and reproducibility of the proposed CNN architecture. The median values for Accuracy (0.969) and F1-score (0.968) lie close to the upper bounds of the distributions, indicating that most individual models perform near the optimum. Notably, the ensemble (red diamond) achieves Accuracy = 0.9757 and F1 = 0.9757, consistently exceeding the maximum of the individual distributions. This confirms that the ensemble strategy not only smooths run-to-run variability but also yields a systematic performance gain over any single model, a key advantage for reliable protein sequence classification.

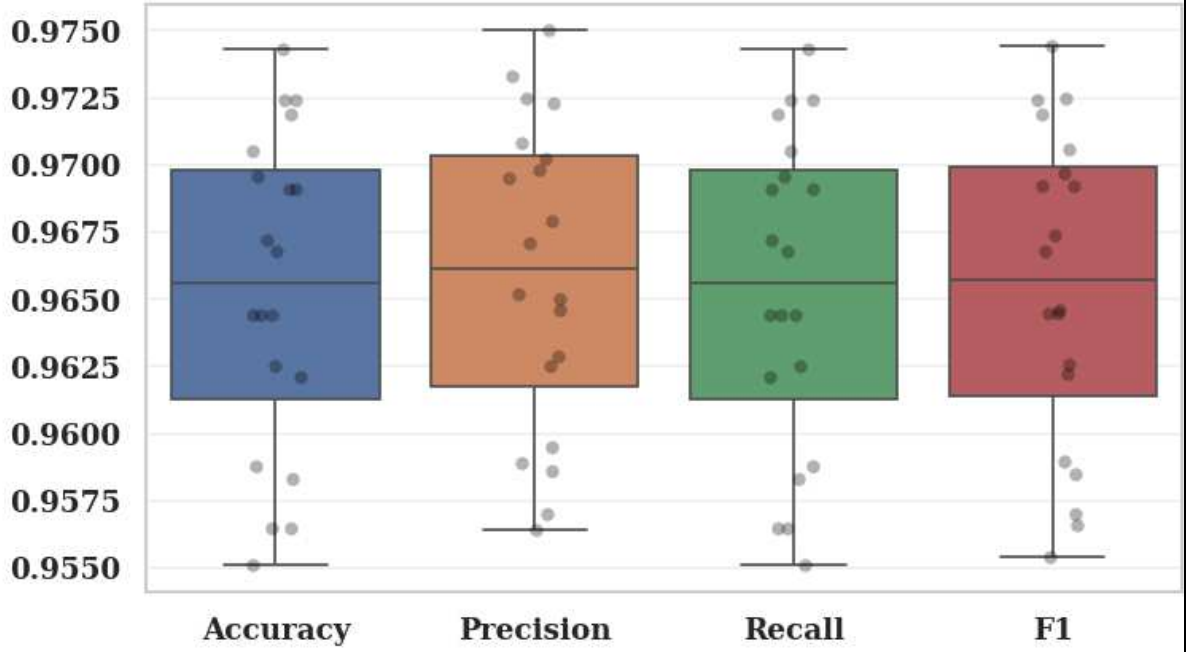


Figure 10. *Distribution of four performance metrics across 20 individual CNN models.*

As shown in Figure 11, the ensemble approach combining the 20 independently trained models substantially improves predictive precision. While individual model performances exhibit some variability (Accuracy ranging from 0.955 to 0.974, F1-score from 0.955 to 0.974), the majority-voting aggregation smooths these fluctuations and raises the ensemble metrics to Accuracy = 0.9757 and F1 = 0.9757, corresponding to relative gains of +0.14% and +0.13% over the best single model (Model_17). This ensemble strategy provides better generalisation when facing ambiguous or noisy samples, as individual errors tend to cancel out. Moreover, the low dispersion of predictions across models (F1 standard deviation < 0.005) demonstrates the stability of the proposed CNN architecture. Therefore, adopting an ensemble of 20 CNNs is a sound methodological choice, ensuring both high performance and enhanced reliability for protein sequence classification.

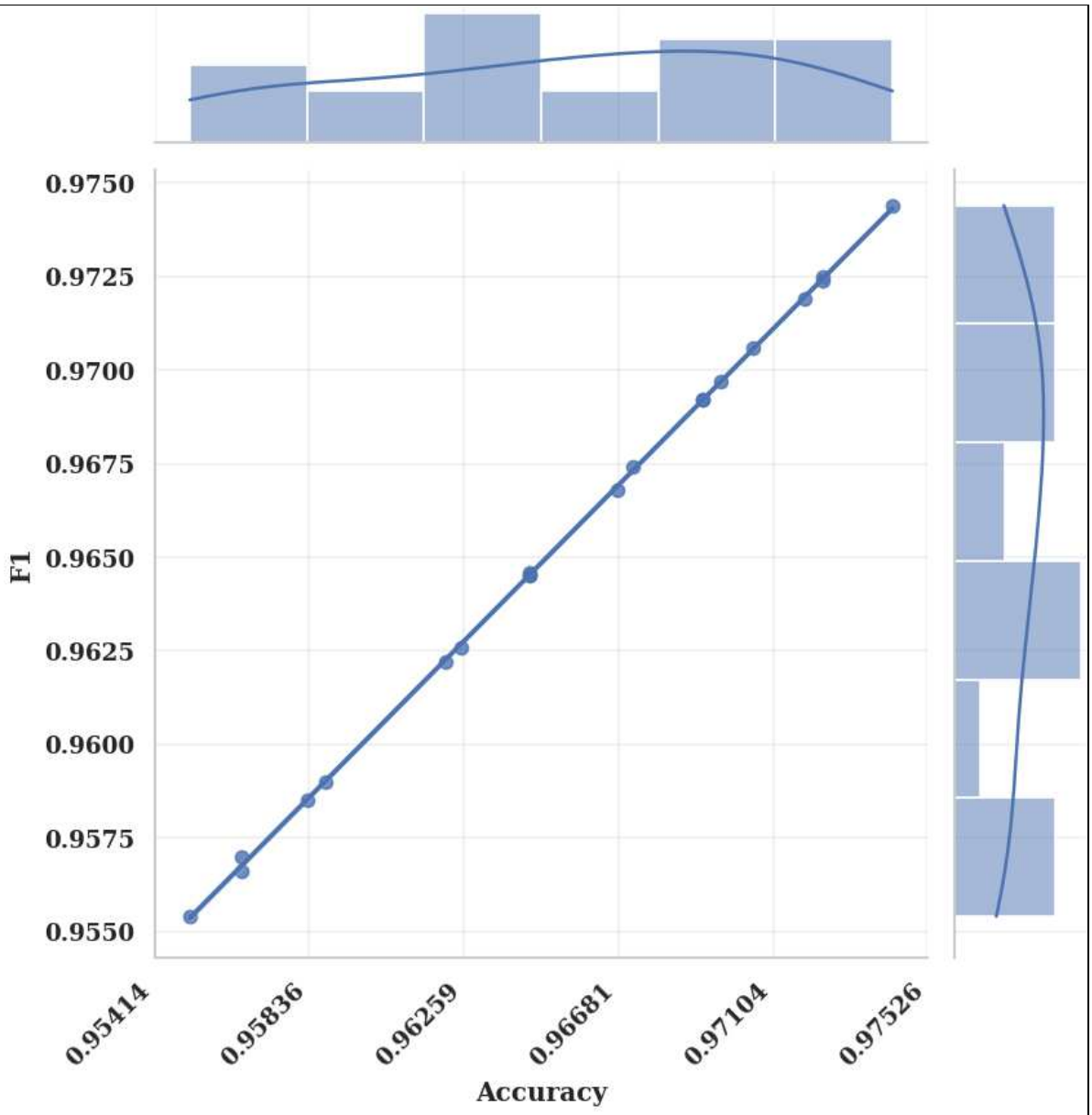


Figure 11. *Accuracy–F1 score regression for the 20 independently trained CNN models.*

DISCUSSION

Male Male infertility continues to present significant challenges in clinical practice, with a substantial proportion of cases remaining idiopathic despite advances in diagnostic methods [1]. Conventional approaches relying primarily on semen analysis according to World Health Organization standards [2] have demonstrated limitations in both sensitivity and mechanistic insight, often failing to provide adequate understanding of the underlying pathophysiology [3]. Our study addresses this critical gap through the development of MIDLPred, an innovative deep learning framework that enables classification of male infertility factors directly from protein sequences.

The implementation of a 1D Convolutional Neural Network architecture combined with an ensemble learning strategy has yielded exceptional performance, achieving a weighted F1-score of 0.96 across four clinically meaningful categories. This methodology represents a significant departure from traditional diagnostic approaches by leveraging the functional information inherent in proteomic data, thereby offering a more fundamental molecular perspective on infertility etiology.

The evolving landscape of artificial intelligence applications in andrology provides important context for evaluating our findings. According to a comprehensive systematic review by Arroyo and colleagues [18], machine learning models for predicting male infertility have demonstrated a median accuracy of 88% across 43 studies, with only seven studies utilizing artificial neural networks specifically and reporting a median accuracy of 84%.

The performance achieved by MIDLPred, with an accuracy of 96%, not only exceeds these established benchmarks but also establishes new standards for computational analysis in male infertility. Whereas existing models predominantly rely on standard clinical parameters or image-based analyses, our approach fundamentally differs through its direct utilization of primary protein sequences. This paradigm shift from analyzing descriptive phenotypic outcomes to deciphering underlying molecular drivers of infertility encoded within the proteome aligns with emerging applications of artificial intelligence in proteomics for biomarker discovery, as highlighted by Mann and colleagues [19].

The biological plausibility of our findings is strengthened when considered alongside contemporary omics research in male infertility. Omolaoye and colleagues [20] have compellingly demonstrated the power of omics technologies in unraveling the complex pathophysiology of this condition. Their transcriptomic analysis identified a core set of consistently dysregulated genes across various infertility conditions, with eight genes LDHC, PDHA2, TNP1, TNP2, ODF1, ODF2, SPINK2, and PCDHB3 emerging as commonly differentially expressed across all disease groups. The protein products of these genes participate in fundamental biological processes including sperm chromatin condensation, flagellar structure, and energy metabolism.

The remarkable accuracy of our sequence-based model suggests an inherent capacity to detect features associated with the functional integrity of these critical proteins and pathways. This

convergence between independent transcriptomic evidence and our proteomic predictions substantially reinforces the biological interpretability of MIDLPred while underscoring the central importance of shared molecular pathways in male infertility.

The model's proficiency in classifying diverse phenotypes aligns well with the complex nature of reproductive dysfunction, wherein genetic factors frequently influence multiple phenotypic outcomes. The identification of 56 genes common between different forms of azoospermia by Omolaoye and colleagues [20] further illustrates this pleiotropic genetic architecture. Our model's performance metrics across categories demonstrate consistent excellence, with precision scores of 98% for both MORPH and MOTILITY categories, 94% for PHYSIO, and 94% for OTHER categories. Similarly, recall values reached 99% for MORPH, 94% for MOTILITY, 97% for OTHER, and 93% for PHYSIO. These results, derived from a substantial sample support of 704, 438, 504, and 490 sequences respectively, highlight the model's robust classification capabilities across diverse infertility manifestations.

The initial chromosomal mapping of infertility-associated genes from the Male Infertility Knowledgebase [10] suggests recurrent genomic hotspots, implying that our model may effectively learn sequence patterns correlating with these enriched biological contexts. This positions MIDLPred not merely as a diagnostic tool but also as a potential hypothesis-generator for identifying novel protein motifs or functional domains critical for male fertility.

The advantages of our sequential approach extend to modeling complex biological relationships without requiring manual feature annotation. MIDLPred autonomously learns hierarchical representations from amino acid sequences, potentially capturing structural and functional motifs that might evade conventional methodologies. This capability proves particularly valuable in biomedical domains characterized by nonlinear relationships and complex interactions [6]. The ensemble strategy employing ten independently trained models further enhances predictive reliability by mitigating individual model variances, as evidenced by improved micro-average AUC scores and reduced misclassification rates across all categories.

Notwithstanding these promising results, several limitations warrant consideration and guide future research directions. The interpretability challenge common to deep learning models represents a primary constraint. Subsequent investigations should incorporate explainable artificial intelligence techniques to elucidate critical residues or domains within protein sequences that drive classification decisions, thereby enhancing model transparency and biological actionability.

Furthermore, while our dataset was meticulously curated from the Male Infertility Knowledgebase [10], expansion to include clinically annotated variants from ethnically diverse populations remains essential for improving model generalizability and clinical applicability. Genetic variability across populations may influence infertility-associated sequence patterns, and a more diverse dataset would better capture this heterogeneity.

The most promising future direction involves developing an integrated multi-omics artificial intelligence framework. Combining the proteomic strengths of MIDLPred with models trained on transcriptomic data [20], genomic variants, and metabolomic profiles would provide a holistic, systems-level perspective on infertility. Such a multimodal approach holds potential for stratifying idiopathic cases into distinct molecular subtypes, thereby paving the way for truly personalized therapeutic strategies in clinical andrology [8].

The clinical implications of MIDLPred are substantial and far-reaching. The model offers potential for more precise molecular diagnosis that could complement routine semen analyses. By categorizing patients into distinct molecular subtypes, MIDLPred could guide more targeted and personalized therapeutic interventions. For instance, patients classified within the MOTILITY category might benefit from therapies specifically addressing flagellar defects, while those in the PHYSIO category might require alternative approaches focusing on underlying testicular dysfunction.

The model's capacity to analyze protein sequences directly also makes it particularly suitable for interpreting newly discovered genetic variants. As next-generation sequencing becomes increasingly prevalent in infertility diagnostics, clinicians often face uncertainty regarding the functional impact of rare or uncharacterized variants [4]. MIDLPred could potentially be adapted to predict phenotypic consequences of such variants, thereby informing genetic counseling and management decisions.

In conclusion, MIDLPred represents a significant advancement in applying deep learning to molecular andrology. By demonstrating accurate classification of infertility factors directly from protein sequences, it addresses important limitations in current diagnostic tools. The convergence of its predictions with existing transcriptomic evidence reinforces biological validity, while its robust architecture and ensemble approach ensure reliability.

With continued development focused on interpretability, data diversity, and multi-omics integration, MIDLPred possesses transformative potential for diagnosing and managing male infertility, ultimately advancing the field toward precision medicine in male reproduction [7,8].

CONCLUSIONS

This study presents MIDLPred, an ensemble learning framework developed for phenotype-oriented classification of male infertility candidate proteins using primary amino acid sequences. By integrating large-scale protein sequence analysis with ensemble learning, the proposed framework achieved strong predictive performance while preserving biological relevance and interpretability.

Unlike conventional diagnostic approaches primarily based on semen analysis and clinical observations, MIDLPred explores the molecular information encoded within protein sequences to identify phenotype-associated signatures linked to male infertility. The results support the hypothesis that infertility-associated proteins contain biologically meaningful sequence patterns that can be

captured through deep learning approaches and used to discriminate among major infertility phenotype categories.

Importantly, MIDLPred should not be considered a standalone diagnostic system. Rather, its primary contribution lies in assisting clinicians, researchers, reproductive biologists, geneticists, and andrologists in prioritizing candidate proteins potentially involved in male infertility mechanisms and directing subsequent experimental investigations. By reducing the complexity of highly heterogeneous infertility phenotypes into biologically coherent groups, the framework provides a practical strategy for exploring molecular associations that may otherwise remain difficult to identify.

Beyond its predictive performance, MIDLPred highlights the potential of artificial intelligence as a complementary tool for molecular andrology, reproductive genetics, and precision medicine. The convergence of our findings with current biological knowledge suggests that protein sequence information may contribute to the identification of candidate biomarkers, facilitate hypothesis generation, and support the interpretation of newly discovered infertility-associated genes and proteins.

The availability of MIDLPred as a user-friendly prediction platform further promotes its adoption by the research community and may accelerate translational studies aimed at understanding the molecular basis of male infertility. Furthermore, the proposed framework may be extended to other areas of reproductive biology and molecular medicine involving complex genotype–phenotype relationships.

Future developments integrating homology-aware validation strategies, explainable artificial intelligence techniques, protein language models, and multi-omics data may further improve the biological interpretability, precision, and clinical utility of the framework. Collectively, these advances may contribute to the development of next-generation computational tools supporting precision andrology and personalized reproductive healthcare.

DECLARATIONS

Clinical Trial Number:

Not applicable.

Consent to Participate declaration:

Not applicable.

Ethics declaration:

Not applicable.

This study was conducted using publicly available protein sequence data and associated annotations obtained from biological databases. No human participants, clinical samples, or animal subjects were directly involved in this research.

Consent for Publication

Not applicable.

Funding

No, this research did not receive funding.

Acknowledgments

The authors acknowledge the support and research environment provided by the Microbiological Engineering and Applications Laboratory (GMA) and the Bioinformatics and Artificial Intelligence Research Team (B2IA), Department of Applied Biology, Faculty of Natural and Life Sciences, University of Constantine 1, Algeria. The authors also acknowledge the support of the LIRE Laboratory, University of Constantine 2, Algeria, for providing an excellent scientific environment that contributed to this research.

We gratefully acknowledge Joseph and Mahale (Joseph & Mahale, 2020) for the development of the Male Infertility Knowledgebase (MIK : https://mik.bicnirrh.res.in/ [PMID: 34363073]), which served as a valuable resource for accessing curated gene and disease associations related to male infertility. The dataset obtained from this resource was instrumental in training and evaluating our deep learning model aimed at predicting male infertility factors from protein sequences.

Competing Interests

The authors declare no conflict of interest.

Author Contributions

A.H.B. conceived the study, curated the dataset, performed the bioinformatics analyses, developed the deep learning models and the MIDLPred web server, analyzed the results, and wrote the main manuscript text. **M.B.** contributed to the study design, supervised the research, and provided expertise in reproductive biology and andrology. **H.C.** contributed to the bioinformatics analyses, artificial intelligence and machine learning methodologies, and co-supervised the project. **S.E.A.** participated in the bioinformatics analyses and data interpretation. **M.A.H**. contributed to the bioinformatics interpretation of the findings and supervised the biological aspects of the study. **A.H.B**. prepared all figures and tables. All authors reviewed, revised, and approved the final manuscript.

Data Availability

The biological data analyzed during this study were obtained from publicly available resources, including the Male Infertility Knowledgebase (MIK: https://mik.bicnirrh.res.in/) and UniProt (https://www.uniprot.org/). The complete source code of MIDLPred is openly available at https://github.com/AHB0738/MIDLPRED, and the web-based prediction platform is accessible at https://midlpred-amob.onrender.com/. The curated dataset generated and analyzed during the current study is available from the corresponding author upon reasonable request. In addition, the

593 curated datasets used in this study are provided as supplementary materials in both Excel (.xlsx) and
594 CSV (.csv) formats to facilitate reproducibility and future research.

REFERENCES

1. Agarwal A, Mulgund A, Hamada A, Chyatte MR. A unique view on male infertility around the globe. *Reprod Biol Endocrinol*. 2015;13(1). doi:10.1186/s12958-015-0032-1

2. World Health Organization. *WHO Laboratory Manual for the Examination and Processing of Human Semen*. 6th ed. World Health Organization; 2021.

3. Carrell DT, De Jonge CJ. The troubling state of the semen analysis. *Andrology*. 2016;4(5):761-762. doi:10.1111/andr.12257

4. Gatimel N, Barbotin A, Mery L, et al. Recommendations for Human Sperm Morphology Assessment in 2025: An Expert Review From the French BLEFCO Group. *Andrology*. Published online November 3, 2025:andr.70134. doi:10.1111/andr.70134

5. Kherraf ZE, Cazin C, Bouker A, et al. Whole-exome sequencing improves the diagnosis and care of men with non-obstructive azoospermia. *Am J Hum Genet*. 2022;109(3):508-517. doi:10.1016/j.ajhg.2022.01.011

6. Ahmed M, Pathan ASK. Deep learning for collective anomaly detection. *Int J Comput Sci Eng*. 2020;21(1):137. doi:10.1504/ijcse.2020.105220

7. Jodar M, Soler-Ventura A, Oliva R. Semen proteomics and male infertility. *J Proteomics*. 2017;162:125-134. doi:10.1016/j.jprot.2016.08.018

8. Venishetty N, Alkassis M, Raheem O. The Role of Artificial Intelligence in Male Infertility: Evaluation and Treatment: A Narrative Review. *Uro*. 2024;4(2):23-35. doi:10.3390/uro4020003

9. Lombardi G, Seoane B, Carbone A. LoRA-DR-suite: adapted embeddings predict intrinsic and soft disorder from protein sequences. *Bioinformatics*. 2025;41(Supplement_1):i439-i448. doi:10.1093/bioinformatics/btaf185

10. Joseph S, Mahale SD. Male Infertility Knowledgebase: decoding the genetic and disease landscape. 2021;2021. doi:DOI: https://doi.org/10.1093/database/baab049

11. Bieniek JM, Drabovich AP, Lo KC. Seminal biomarkers for the evaluation of male infertility. *Asian J Androl*. 2016;18(3):426-433. doi:10.4103/1008-682X.175781

12. Jia X, Yin Z, Peng Y. Gene differential co-expression analysis of male infertility patients based on statistical and machine learning methods. *Front Microbiol*. 2023;14. doi:10.3389/fmicb.2023.1092143

13. Homanen R, McBride N, Hudson N. Artificial intelligence and assisted reproductive technology: Applying a reproductive justice lens. *Eur J Womens Stud*. 2024;31(2):262-276. doi:10.1177/13505068241258053

14. Ponting CP. A novel domain suggests a ciliary function for *ASPM* , a brain size determining gene. *Bioinformatics*. 2006;22(9):1031-1035. doi:10.1093/bioinformatics/btl022

15. Yang R, Stendahl AM, Vigh-Conrad KA, Held M, Lima AC, Conrad DF. SATINN: an automated neural network-based classification of testicular sections allows for high-throughput histopathology of mouse mutants. Peng H, ed. *Bioinformatics*. 2022;38(23):5288-5298. doi:10.1093/bioinformatics/btac673

16. Kobayashi H. Potential for artificial intelligence in medicine and its application to male infertility. *Reprod Med Biol*. 2024;23(1):e12590. doi:10.1002/rmb2.12590

17. Liang B, Wang M. Deep learning-based approach for sperm morphology analysis. *BMC Urol*. 2025;25(1):261. doi:10.1186/s12894-025-01946-w

18. Schmeis Arroyo V, Iosa M, Antonucci G, De Bartolo D. Predicting Male Infertility Using Artificial Neural Networks: A Review of the Literature. *Healthcare*. 2024;12(7):781. doi:10.3390/healthcare12070781

19. Mann M, Kumar C, Zeng WF, Strauss MT. Artificial intelligence for proteomics and biomarker discovery. *Cell Syst*. 2021;12(8):759-770. doi:10.1016/j.cels.2021.06.006

20. Omolaoye TS, Omolaoye VA, Kandasamy RK, Hachim MY, Du Plessis SS. Omics and Male Infertility: Highlighting the Application of Transcriptomic Data. *Life*. 2022;12(2):280. doi:10.3390/life12020280

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- MIDLPredDATASET2026.xlsx