**Development of a Machine Learning Model Based on Neural Networks for Assessing the Probability of Pregnancy in IVF Protocols.**

**Sergeev S.A. Diakova Y.B.**

**Introduction:**

In recent decades, neural networks and machine learning models have become key tools in various fields, including computer vision, natural language processing, recommender systems, medical diagnostics and in the field of assisted reproductive technologies. They serve as the foundation for creating algorithms capable of extracting complex dependencies from data, making predictions, and making decisions based on these dependencies. In some cases, neural networks can identify a broader spectrum of associations than other statistical methods, thanks to their ability to recognize highly nonlinear associations among input parameters.

Deep Learning Neural Networks (DNN) represent a powerful tool for data analysis, forecasting, and process automation. Their effectiveness depends on data quality, network architecture selection, and training process optimization. Neural networks consist of interconnected nodes called neurons, organized into layers: input, hidden, and output. The input layer receives data, hidden layers perform computations, and the output layer provides results or predictions. Each neuron is linked to neurons in the next layer through weights, determining the strength of the connection between neurons. Thus, the neural network finds a correct method of mathematical transformations to convert input data into output, regardless of their linear or nonlinear correlation.

According to international data, the overall clinical probability of pregnancy after embryo transfer is approximately 30%, and this probability varies significantly in different clinics, depending on input parameters that are beyond our control. It also depends on the laboratory's activities, characterized by its quality indicators - Key Performance Indicators (KPI) - applied in the quality management system for minimizing and controlling risks.

**Dataset description:**

|  | **Count** | **Mean** | **Std** | **Min** | **25%** | **50%** | **75%** | **Max** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Age | 3856 | 35.52 | 5.26 | 19 | 32 | 36 | 39 | 56 |
| Attempt | 3856 | 2.1 | 1.91 | 0 | 1 | 1 | 3 | 36 |
| Follicles | 3856 | 8.86 | 5.95 | 0 | 4 | 8 | 12 | 60 |
| OCC | 3856 | 7.46 | 4.96 | 0 | 4 | 7 | 10 | 46 |
| Fertilized | 3856 | 5.74 | 3.96 | 1 | 3 | 5 | 8 | 40 |
| 2 pN | 3856 | 4.03 | 2.98 | 0 | 2 | 3 | 6 | 23 |
| Cleav. D3 | 3856 | 3.95 | 2.92 | 0 | 2 | 3 | 5 | 23 |
| All Bl | 3856 | 1.89 | 2.5 | 0 | 0 | 1 | 3 | 17 |
| Good Bl | 3856 | 0.89 | 1.53 | 0 | 0 | 0 | 1 | 11 |
| FR | 3856 | 0.73 | 0.25 | 0 | 0.56 | 0.75 | 1 | 2 |
| CR | 3856 | 0.96 | 0.16 | 0 | 1 | 1 | 1 | 1 |
| BlR | 3856 | 0.33 | 0.36 | 0 | 0 | 0.25 | 0.67 | 1 |
| GBR | 3856 | 0.16 | 0.25 | 0 | 0 | 0 | 0.29 | 1 |
| OCCR | 3856 | 0.85 | 0.28 | 0 | 0.75 | 0.9 | 1 | 14 |
| D5 embryos | 3856 | 3.17 | 3.51 | 0 | 0 | 2 | 5 | 31 |
| Cryo | 3856 | 1.05 | 1.53 | 0 | 0 | 0 | 2 | 13 |
| ETr Day | 3856 | 3.59 | 1.13 | 0 | 3 | 3 | 5 | 7 |
| ET Embryos | 3856 | 1.45 | 0.58 | 1 | 1 | 1 | 2 | 4 |
| Pregnancy | 3856 | 0.38 | 0.49 | 0 | 0 | 0 | 1 | 1 |
| KPIScore | 3856 | 14.93 | 4.62 | 5 | 11 | 15 | 19 | 25 |

**Model Parameter Selection:**

The standard prediction model (Kaufmann et al., 1997. DOI: 10.1093/humrep/12.7.1454) based on the following data: Age, Attempt number, Number of follicles, Number of oocytes retrieved, Number of inseminated oocytes, 2PN, Number of cleaved embryos, Number of blastocysts, was simulated in this study:

Epoch 1/10

106/106 [==============================] - 1s 4ms/step - loss: 0.6894 - accuracy: 0.5117 - val\_loss: 0.6794 - val\_accuracy: 0.5498

Epoch 2/10

106/106 [==============================] - 0s 3ms/step - loss: 0.6695 - accuracy: 0.5568 - val\_loss: 0.6696 - val\_accuracy: 0.5486

Epoch 3/10

106/106 [==============================] - 0s 3ms/step - loss: 0.6622 - accuracy: 0.5666 - val\_loss: 0.6647 - val\_accuracy: 0.5498

Epoch 4/10

106/106 [==============================] - 0s 3ms/step - loss: 0.6602 - accuracy: 0.5645 - val\_loss: 0.6644 - val\_accuracy: 0.5628

Epoch 5/10

106/106 [==============================] - 0s 3ms/step - loss: 0.6567 - accuracy: 0.5642 - val\_loss: 0.6642 - val\_accuracy: 0.5557

Epoch 6/10

106/106 [==============================] - 0s 2ms/step - loss: 0.6549 - accuracy: 0.5686 - val\_loss: 0.6647 - val\_accuracy: 0.5746

Epoch 7/10

106/106 [==============================] - 0s 3ms/step - loss: 0.6532 - accuracy: 0.5689 - val\_loss: 0.6659 - val\_accuracy: 0.5664

Epoch 8/10

106/106 [==============================] - 0s 3ms/step - loss: 0.6520 - accuracy: 0.5710 - val\_loss: 0.6659 - val\_accuracy: 0.5711

Epoch 9/10

106/106 [==============================] - 0s 2ms/step - loss: 0.6510 - accuracy: 0.5743 - val\_loss: 0.6650 - val\_accuracy: 0.5723

Epoch 10/10

106/106 [==============================] - 0s 3ms/step - loss: 0.6506 - accuracy: 0.5669 - val\_loss: 0.6658 - val\_accuracy: 0.5735

27/27 [==============================] - 0s 2ms/step - loss: 0.6658 - accuracy: 0.5735

Unfortunately, its accuracy and reproducibility are insufficient for predicting the likelihood of clinical pregnancy occurrence in the IVF protocol (loss: 0.6658 - accuracy: 0.5735). This led to the analysis of additional publications and studies involving the modeling of raw protocol data and its embryological component in the prediction of outcomes based on laboratory input parameters (Liu et al., 2023; Blank et al., 2019; Raef et al., 2020; Bormann et al., 2020; Hariton et al., 2021). The aim was to select optimal input parameters for predicting the outcome of embryo transfer.

Our deep learning neural network model additionally incorporates data from the KPI of the Vienna Consensus, which shows a correlational dependence with the occurrence of pregnancy. It also includes the rankings of the sum of laboratory and clinical parameters (KPI Score), used to forecast the chance of pregnancy occurrence (Franco et al., 2017). The developed neural network makes predictions based on 20 parameters (14 recorded in the laboratory database and available in the MedWork report, and 6 mathematically calculated, including the predictive KPIScore for implantation):

"Age", "Attempt number", "Number of follicles", "Number of oocytes retrieved", "Number of inseminated oocytes", "2PN", "Number of cleaved embryos on day 3", "Number of blastocysts", "Number of high-quality blastocysts", "Fertilization rate", "Cleavage rate", "Blastocyst formation rate", "High-quality blastocyst formation rate", "Oocyte retrieval rate", "Number of embryos on day 5", "Frozen embryos", "Transfer day", "Embryos transferred", "KPIScore".

Using machine learning algorithms, the values of their correlation with pregnancy occurrence for selected parameters, considering their significance in predicting the outcome of embryo transfer, were analyzed. The following weights were obtained:

| **Feature** | **Importance** |
| --- | --- |
| Age | 0.175163 |
| Frequency of high-quality blastocyst formation | 0.091709 |
| KPIScore | 0.082998 |
| Number of follicles | 0.081302 |
| Number of MII oocytes | 0.068844 |
| Fertilization rate | 0.062863 |
| Frequency of obtaining MII oocytes | 0.054834 |
| Frequency of blastocyst formation | 0.050492 |
| Attempt number | 0.050036 |
| Number of inseminated oocytes | 0.047903 |
| Transfer day | 0.038013 |
| Frozen embryos | 0.033613 |
| Number of 5-day embryos | 0.028328 |
| Frequency of cleavage | 0.027952 |
| Number of embryos transferred | 0.027678 |
| Number of high-quality blastocysts transferred | 0.027531 |
| 2 pN | 0.019465 |
| Number of embryos dividing on day 3 | 0.017543 |
| Number of blastocysts | 0.013733 |

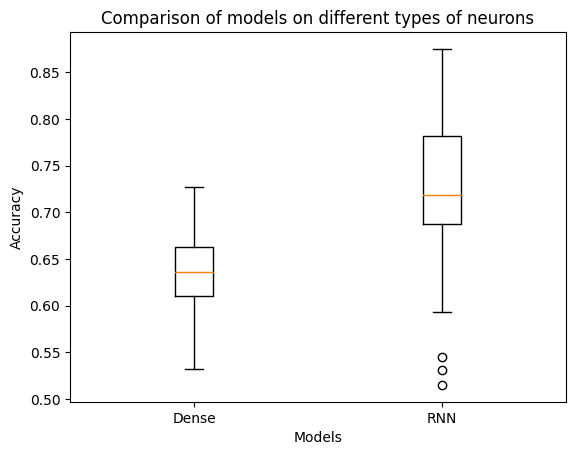
**Model Compilation:**

The neural network model was developed and executed in the CPU Google Colab environment using Python 3.5 programming language. The implementation utilized the Keras library, version 2.14.0.

For the first model we take neural network architecture from published working models (Goyal et al., 2020). The architecture contains a total of 9 dense layers, each neuron output values In the first half of the model, neurons in each layer get increased precisely two times the previous layer. The second half follows a decreasing rate of two neurons per layer, making the last layer one. Totally 32x64x128x256x512x256x128x64x32x1 neurons were used. That model proposed by the authors was fed with data from the Human Fertilisation and Embryology Authority (HFEA), comprising 25 clinical parameters, which, nevertheless, proved insufficient for describing the processes of individual embryo development in the laboratory. Despite achieving good results in predicting outcomes in our data (TEST Accuracy: 0.6749), the model proved to be highly complex and required significant optimization methods to prevent overfitting. Adapting the model to our dataset posed a challenging task.

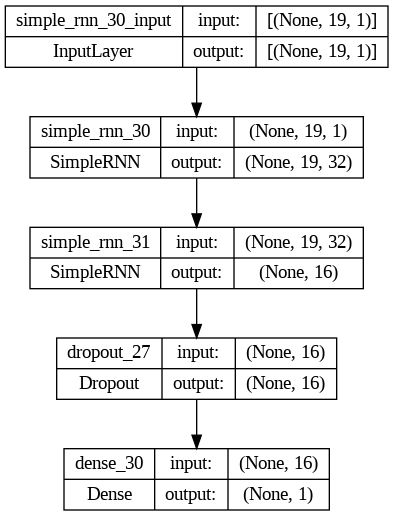
The proposed neural network model is a sequential model consisting of two layers of SimpleRNN recurrent neurons and one layer of a dense neural network (Dense) with sigmoid activation, designed for solving a classification task. The model starts with an input layer that takes data with a dimension corresponding to the number of selected features. Then, hidden layers are added sequentially, each containing neurons with ReLU (Rectified Linear Activation) activation function, in addition to the final layer with a sigmoid activation function. The model structure was developed based on existing publications in the field of artificial neural networks in ART (Vogiatzi et al., 2019. doi: 10.1007/s10815-019-01498-7). However, in all these works, the authors considered input parameters as independent variables, which is unfortunately not applicable in our model. When using such models with sequentially connected neurons, their prediction accuracy does not exceed 63% (loss: 0.6797 - accuracy: 0.5764). Since KPI data and the final score prediction (KPIScore) are related to the model's input parameters, a recurrent type of RNN neurons was chosen for building the deep learning neural network.

The RNN neurons exhibit higher average accuracy (0.7236) compared to Dense neurons (0.6362). The RNN model shows a wider range of accuracy (0.5152 to 0.875) compared to the Dense model (0.5325 to 0.7273). The standard deviation for accuracy is also higher in the RNN model (0.0764) compared to the Dense model (0.0410). The RNN model achieves a higher maximum accuracy (0.875) compared to the Dense model (0.7273). These differences are statistically significant: t-Statistics: -7.055274039431632 p-value: 2.4699782266522906e-10.



The model architecture includes the following layers:

1. Recurrent Neural Network (RNN) Layer:
2. The first layer, SimpleRNN with 32 neurons and ReLU activation, takes input features.
3. The second layer, SimpleRNN with 16 neurons and ReLU activation.
4. Dense Neural Network (Dense) Layer:
5. The final layer with one neuron and Sigmoid activation, designed for binary classification. Sigmoid is used to predict the probability of a positive class (in this case, a positive outcome of the embryo transfer procedure).



Added Dropout layer: this layer helps prevent model overfitting by randomly excluding neurons during training, thereby improving the model's generalization ability.

L2 and L1 regularization: Applying L2 and L1 regularization to the corresponding layers of the model to control overfitting by adding penalties to the model's weights allows achieving greater accuracy and reproducibility of prediction results during training.

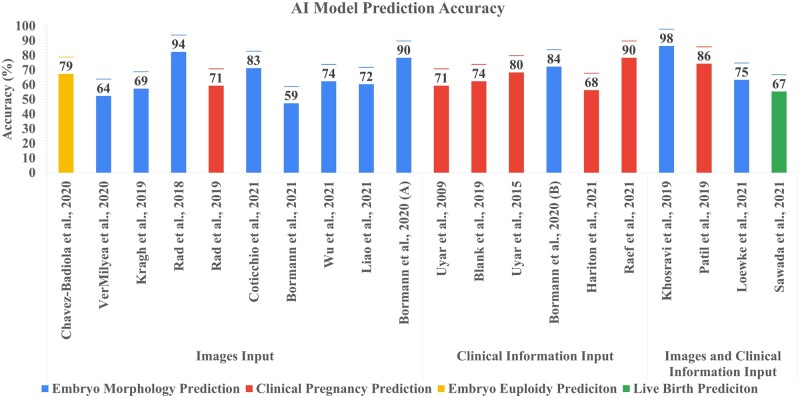
Optimizer and loss function: This model is optimized using the gradient descent method with the Adam optimization algorithm and a specified learning rate of 0.0001. Binary cross-entropy is used as the loss function. The model is trained on a data array for 12 epochs (independently passes through the entire dataframe, splitting it into training and testing parts) with a batch size of 8.

**Model Training:**

The model was trained on the MedWork database over a 10-year period from 2013 to 2023. A total of 7642 protocols were available for analysis during this time. After selecting protocols containing all the necessary information for neural network training, 3858 cases of embryo transfer with known outcomes were used to form the training (80%) and validation (20%) set. The model was developed using the KERAS library and the Python programming language.

After training the model on the presented cycle dataset with decision correction, sufficient accuracy and reproducibility were achieved, comparable to those of artificial intelligence models in time-lapse studies (Bormann et al., 2020. doi: 10.7554/eLife.55301; Louis et al., 2021. doi: 10.1007/s10815-021-02123-2; Salih et al., 2023. doi: 10.1093/hropen/hoad031.):

Validation accuracy: 0.67 - 0.77, SD 0.0354

****

The obtained probability of success in predicting the developed model aligns with existing published systematic reviews and meta-analyses of data in the field of using neural networks in ART (Salih et al., 2023. doi: 10.1093/hropen/hoad031). Similar studies on the use of artificial intelligence models in predicting outcomes of ART programs have an average accuracy of 77.8% (range 68–90%). However, in the modeling process in this study, the average accuracy of such models for our database was 63%.

A comparison was conducted between the developed model and other machine learning approaches described in the literature, specifically those by Lu et al., 2023 (doi: 10.1186/s12884-023-05775-3):

Gradient Boosting Model Accuracy: 0.6244

Random Forest Model Accuracy on Test Data: 0.5816

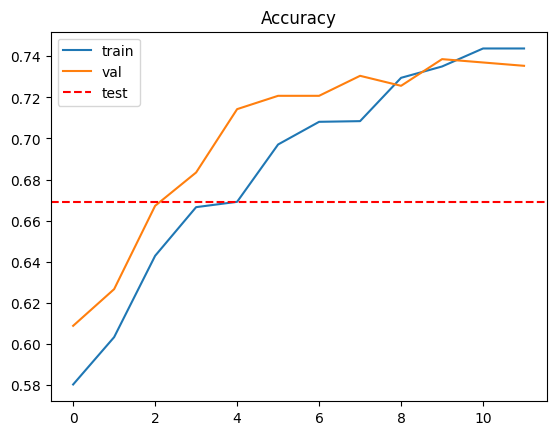
Decision Tree Model Accuracy on Test Data: 0.5479

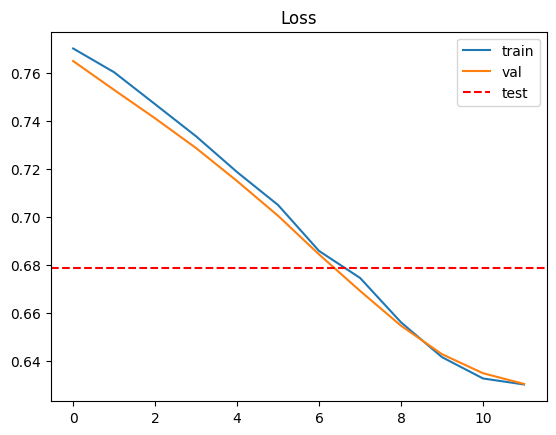
The developed model exhibits a significant higher (U-Statistic: 230.0, p-value: 0.002) accuracy (0.67, AUC-ROC: 0.63; SD = 0.058) compared to the Gradient Boosting, Random Forest, and Decision Tree models on the provided test data. But lower than was published (Blank et al., 2018): Accuracy 0.74 ± 0.03 for the proposed RFM on their datasets that means a high difference between clinic’s data.

For model propria fitting and validation we added 4000 cryo protocols with known implantation data to the previous dataset (total 7858 protocols were used). And split that data to train (70%), validation (20%) and test (10%) set. On this data a high stability of model accuracy was observed:

Accuracy: 0.7438 - val\_accuracy: 0.7370; Sensitivity: 0.6; Specificity: 0.6634615384615384; PPV: 0.46153846153846156; NPV: 0.7752808988764045; FPR: 0.33653846153846156; FNR: 0.4; Overall Accuracy (test): 0.6428571428571429; Odds Ratio: 2.9571428571428573

The graphs presented depict the model training process and validation of prediction values using embryology laboratory KPI data and patient input parameters. The generated graphs demonstrate the correct model development strategy, justified efficiency assessment algorithm, and proper compilation of neural layers in the model.



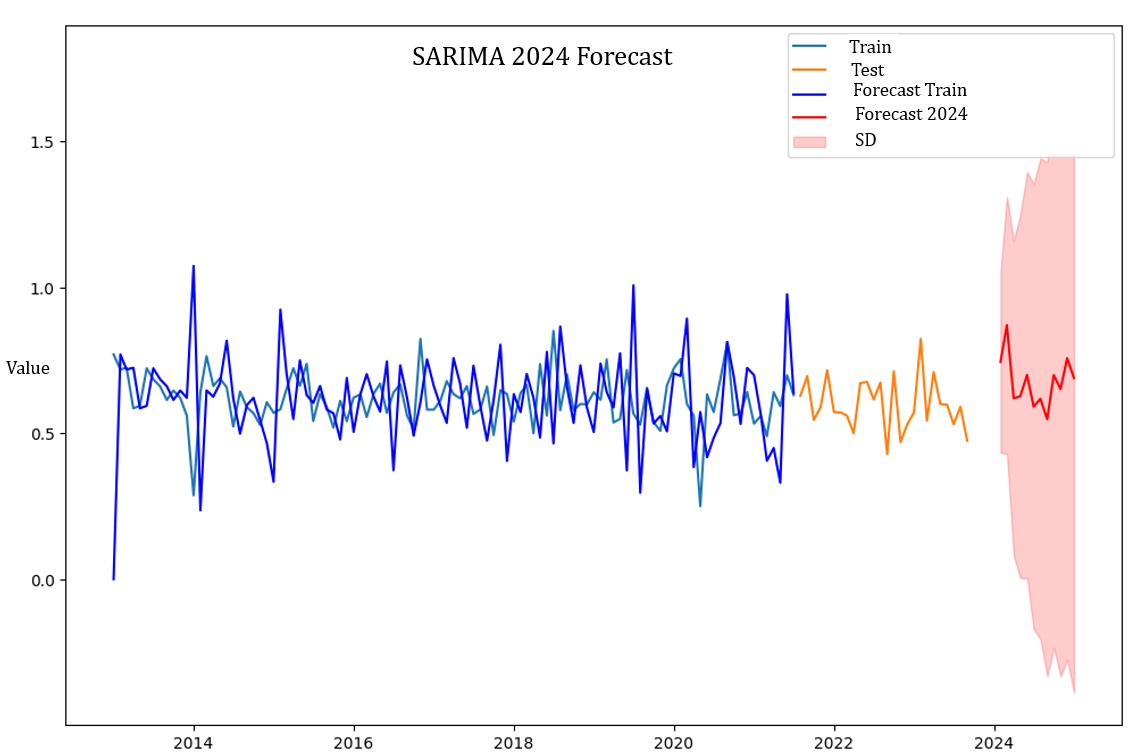


The Model AUC ( 0.68; SD = 0.054) comparison with other AI models in the literature has revealed the advantage of the developed neural network model in terms of time lapses for implantation probability prediction: 0.573 ( Chamayou et., al. 2013); 0.629 (Basile et al., 2015); 0.543 (Dal Canto et al., 2021); 0.638 (Bori et.al., 2022).

During the execution of the model code, the laboratory's embryology KPI is automatically calculated according to the recommendations of the Vienna Consensus. This serves as a convenient tool for internal quality control, allowing for quick response and decision-making regarding changes in the process when deviations from reference values are detected. The report is output in the form of a table "predictions.xlsx," and its data is specific to a particular IVF cycle.

The next use of SARIMA (Seasonal Autoregressive Integrated Moving Average) models provides the opportunity for further forecasting of obtained KPIs (Key Performance Indicators) and subsequent comparison with actual performance metrics. This approach allows for the development of a versatile and effective tool for precise internal quality control in the laboratory based on recorded quality performance indicators. SARIMA models are a powerful time series tool that takes into account trends and seasonal variations in data. Forecasting KPIs using SARIMA enables the laboratory to analyze expected indicators in the future, identify potential risks, and take measures to manage the quality of operations.

Subsequent comparison of forecasted KPIs with actual data forms the basis for a systematic analysis of laboratory efficiency. As new data becomes available, the models can be periodically updated, allowing adjustments to forecasts based on changing conditions and requirements. This approach to internal quality control provides the laboratory with a tool for more predictable process management, early detection of potential issues, and optimization of operations based on data and analytics.



**Validation Analysis of the Model:**

A validation analysis of the model was conducted by running it on 12 randomly shuffled pseudo-randomly generated data frames, subjected to sequential and independent analyses. The model performance on an independent test: specificity in 96.8% of cases, with false-positive predictions occurring in 15.5% of cases and an average ability to correctly predict the occurrence of clinical pregnancy in 63% of cases.

An analysis of the correlation between model accuracy and various age groups of patients (up to 29; 30-34; 35-39; over 40 years) was conducted, considering different methods of fertilization (IVF and ICSI). No significant differences were observed across the specified age groups when employing these diverse fertilization methods. This suggests that the developed model's accuracy remains consistent across age brackets and is not influenced significantly by the age of the patients. Additionally, the lack of observed differences in model performance based on the fertilization methods (IVF and ICSI) indicates that the model is robust and effective irrespective of the specific assisted reproductive technology used.

Validation of the concordance between real and predicted values of pregnancy occurrence using this model was carried out on 1600 IVF protocols at the validation dataset. An analysis was performed on how often the model correctly predicts the probability of clinical pregnancy occurrence – its accuracy is 0.7685 with an 86% sensitivity for this class (the proportion of actual positive results detected by the model out of all true positive results). The summary metric, which takes both accuracy and sensitivity into account (F1-score), is 0.71, comparable to the literature data of similar models (Liu et al., 2023).

Cross-validation for the neural network model was conducted using scikit-learn and Keras to evaluate the model on 50 different data splits (folds). Different combinations of training and test sets were used during cross-validation, providing a more robust and reliable assessment of model performance:

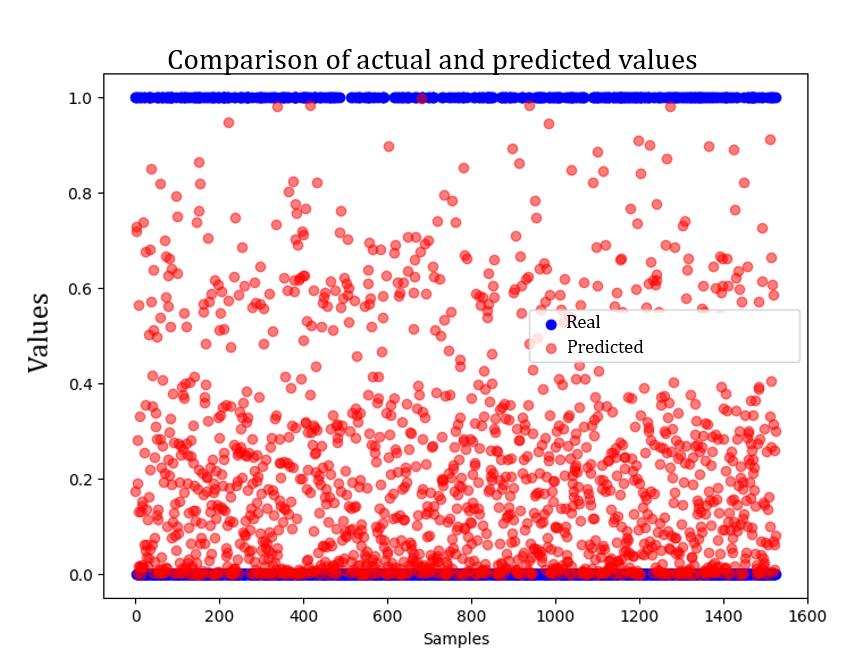
Average accuracy: 0.6789876782894134

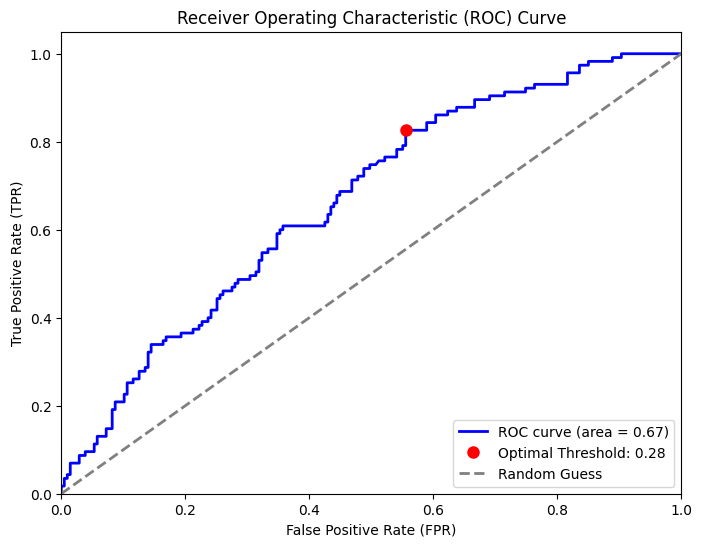
Standard deviation of accuracy: 0.05446925480159457

Maximum accuracy: 0.8311688303947449

For our purposes, binary classification is used – pregnancy either occurred (1) or did not (0). To reduce the likelihood of erroneous predictions biased towards negative outcomes, a decision threshold of 0.46 was set, adjusted based on the confusion matrix so that real events corresponded to the calculated values.

An analysis of the model's erroneous predictions was conducted through ROC analysis. The results were used to create an AUC graph, allowing an assessment of the quality of binary classification. It reflects the relationship between the proportion of clinically pregnant occurrences correctly classified as leading to clinical pregnancy and the proportion of clinically pregnant occurrences incorrectly classified as not bearing such a feature. The ROC curve represents the boundary of random decision-making and forecast inconsistency. In the used model, all prediction data for the occurrence of pregnancy are above this boundary, and the area under the curve is greater than the area under the limiting line. This allows for the use of the final classification of our model as a potential predictive indicator for the occurrence of clinical pregnancy in the IVF protocol.

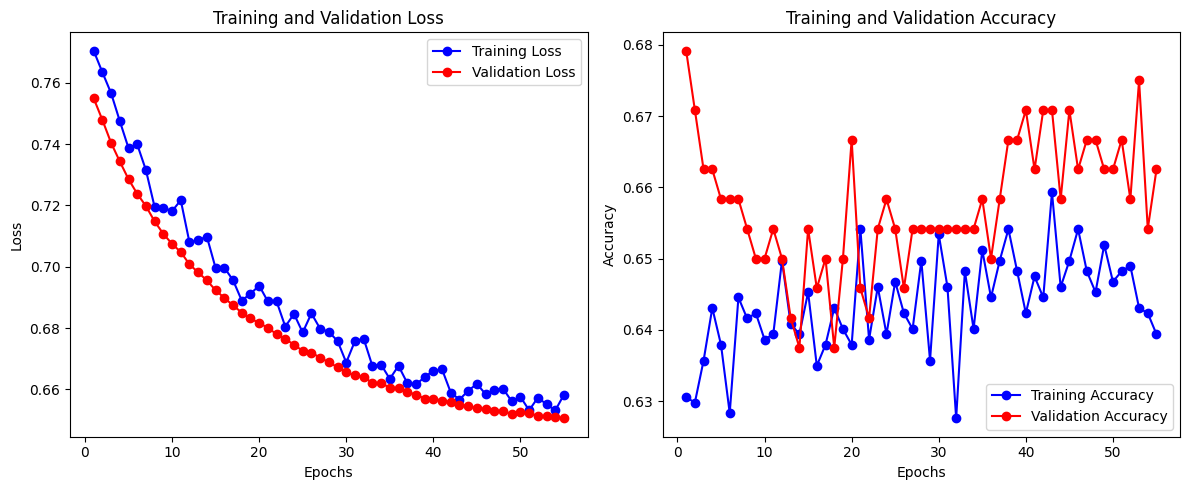


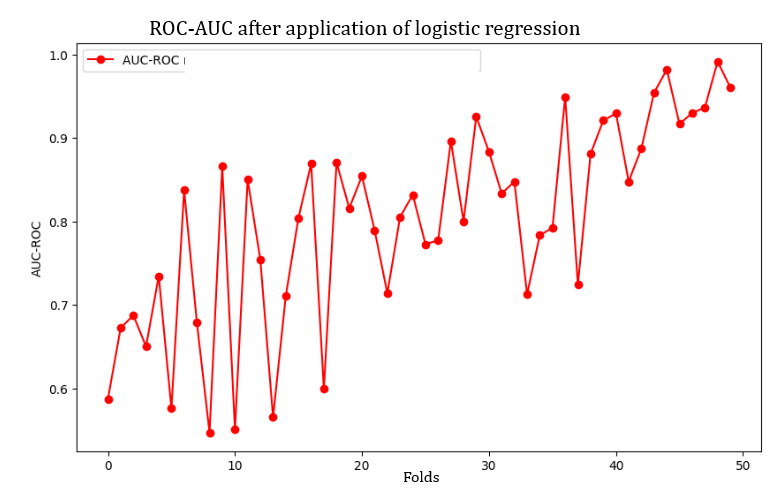


**Model Calibration**

After training a neural network model, we utilize the probabilities obtained from it and apply logistic regression for calibration. Calibration is performed using CalibratedClassifierCV from scikit-learn, which applies logistic regression to align probabilities. Thus, we create a model that takes outputs from the neural network and calibrates them, improving the quality of probability predictions AUC = 0.86; SD = 0.061 was obtained.

In this process, both models play distinct roles: the neural network is employed to extract complex patterns from the data, while logistic regression is applied to calibrate the probabilities, making them more accurate and interpretable. Validation process of that combined model on 50 epochs was used:





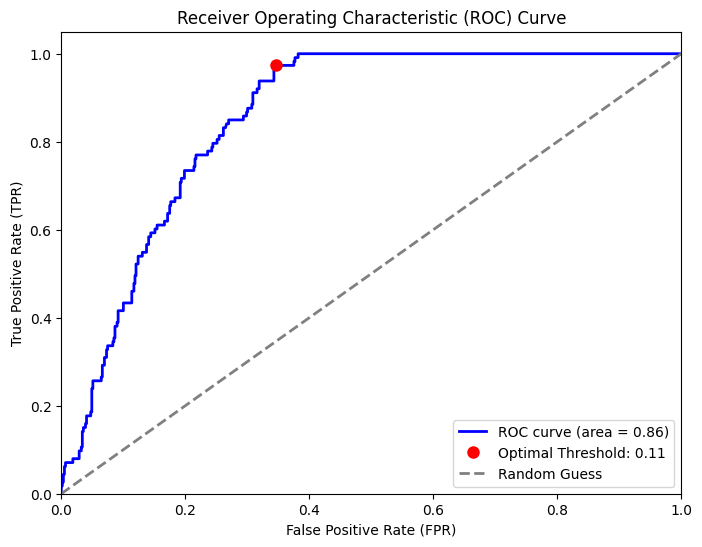
The model was fine-tuned based on the database from 2018 to 2023 (3500 protocols with known implantation data) for all embryos. During this period, a comprehensive quality control system was implemented in the laboratory, and there was no variability in approaches to embryo culture and transfer.

Average Accuracy: 0.8553963792324066

Standard Deviation of Accuracy: 0.027059927998945393

Maximum Accuracy: 0.9142857193946838

Minimum Accuracy: 0.7857142686843872



A comparative analysis of prediction errors was conducted between the developed model and the AI Vitrolife for time-lapse systems based on the characteristics of the ROC-AUC curve. During validation, the obtained model exhibited a similar ROC-AUC curve (U- Statistic: 145.0, p-value: 0.471) with an area of 0.67, comparable to the AI KIDScore™ (for cases of embryo transfer with known implantation data AUC = 0.66, AUC range of 0.60-0.75 for KID embryos across different clinics). Additionally, after fitting the model, it demonstrated a similar curve to the KIDScore™ (0.86 vs 0.89) for all time lapse embryos. This convincingly illustrates the equivalent accuracy in predicting the frequency of pregnancy occurrence for the developed model, with comparable precision to commercially utilized time-lapse systems (Fréour et al.,2015, Tran et al., 2019, Reignier et al., 2019, Berntsen et al., 2022).

Using this model with logistic regression calibration, it is possible to calculate the probability of pregnancy occurrence in the dataset. Corrections to Positive Predictive Value (PPV) and Negative Predictive Value (NPV) were applied to the calculated pregnancy occurrence frequency, determined during the model validation. This approach allows the characterization of the patient population at a specific time and performance analysis based on the correspondence between real results and the calculated pregnancy occurrence value.

The DNN model demonstrates high specificity (86.4%) and an average ability to correctly predict clinical pregnancy (Average accuracy: 78%, Standard deviation of accuracy: 4%, Maximum accuracy: 87.5%). The accuracy of predicting clinical pregnancy and sensitivity in the model compares favorably with the Irvine Scientific Life Whisperer model (VerMilyea et al., 2020): accuracy for predicting clinical pregnancy 76.85%, sensitivity of 70.1% for viable embryos while maintaining a specificity of 60.5%, combined accuracy of 64.3%. And has a better ROC-AUC Score AUC = 0.86 (0.82-0.89) than traditionally used predictive models AUC = 0.6335 (0.6202–0.6367) Nelson SM, Lawlor DA (2011).

| **Metrics:** | **AUC** | **Sensitivity** | **Specificity** | **PPV** | **NPV** | **FPR** | **FNR** | **Accuracy** |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | 0,849 | 0,714 | 0,833 | 0,769 | 0,789 | 0,167 | 0,286 | 0,781 |
|  | 0,888 | 0,583 | 0,850 | 0,700 | 0,773 | 0,150 | 0,417 | 0,750 |
|  | 0,895 | 0,400 | 0,909 | 0,667 | 0,769 | 0,091 | 0,600 | 0,750 |
|  | 0,823 | 0,700 | 0,773 | 0,583 | 0,850 | 0,227 | 0,300 | 0,750 |
|  | 0,864 | 0,700 | 0,955 | 0,875 | 0,875 | 0,045 | 0,300 | 0,875 |
| **Mean** | **86,37%** | **61,95%** | **86,39%** | **71,88%** | **81,13%** | **13,61%** | **38,05%** | **78,13%** |
| **SD** | 0,024 | 0,109 | 0,057 | 0,090 | 0,039 | 0,057 | 0,109 | 0,044 |

The obtained model can be used to analyze real patient data for any time period according to MedWork data. For example, in 2023, the model's output would be represented in the "predictions.xlsx" table, with indicators of a complete match, representing pregnancies contrary to the model's expectation, and indicating cases requiring in-depth analysis of unsuccessful transfers.

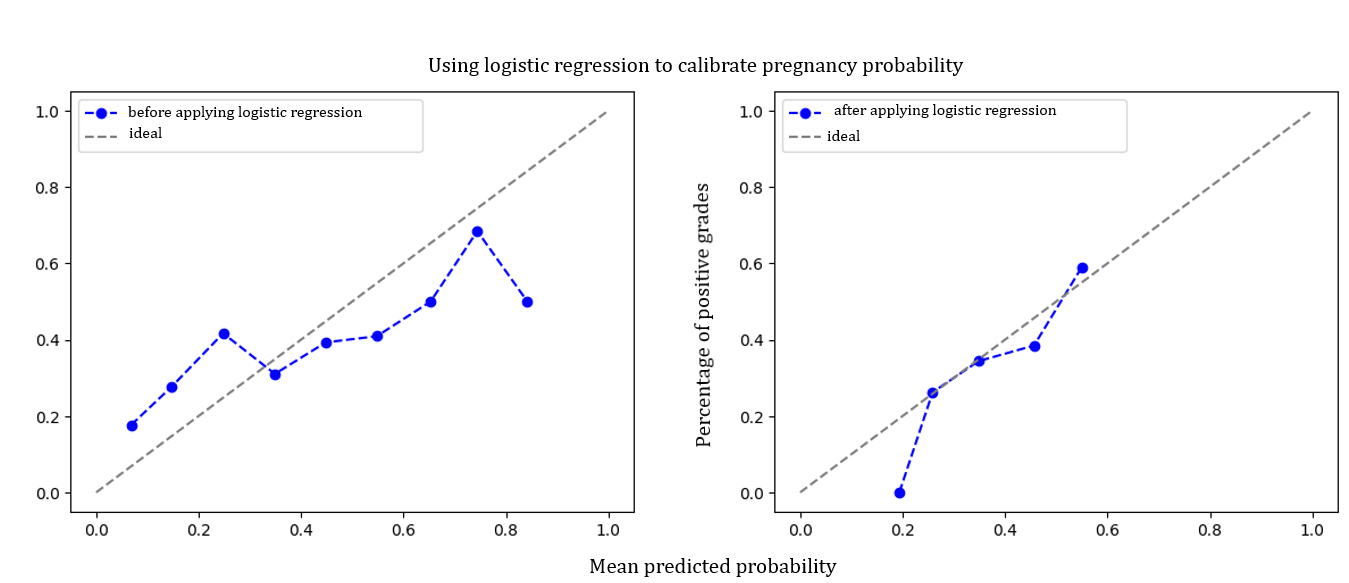
The threshold value for pregnancy occurrence is a cumulative score of 15 in KPIScore, taking into account the confidence interval for implantation (16%-30%). there was a significant difference (p<0.0001) with respect to the total KPIs mean score among the group of patients with clinical pregnancy (total KPIsscore=20.4±3.7) and the group without clinical pregnancy (total KPIs-score=15.9±5). When comparing the predictions of the developed model and KPIScore (A/B model stat test), a 7% difference was observed.

In the test set of 2023 protocols, the successful prediction of the absence of pregnancy occurrence was 86%. An analysis of pregnancy occurrence frequency in protocols with Preimplantation Genetic Testing (PGT) and the transfer of a euploid embryo of good quality revealed a 27% error in predicting clinical pregnancy occurrence AUC = 0.68 (0.75-0.62), Odds Ratio = 6.66.

Accuracy: 0.7850 - val\_accuracy: 0.7685

Specificity: 0.9688796680497925

Those results were higher than in Meseguer group model IDAScore V2 for the same patient category AUC = 0.654 (Bori, L., Meseguer, et al., 2022).



In the process of validating the accuracy of pregnancy occurrence prediction, a quarterly analysis for 2022-2023 was conducted, comparing the model's predictions with the real clinical report and MedWork statistics AUC = 0.86 (0.82-0.89):

1st Quarter 2022: Estimated pregnancy occurrence rate with PPV and NPV: 25.73%, report: 23.51%

2nd Quarter 2022: Estimated pregnancy occurrence rate with PPV and NPV: 31.66%, report: 34.78%

3rd Quarter 2022: Estimated pregnancy occurrence rate with PPV and NPV: 26.48%, report: 24.38%

4th Quarter 2022: Estimated pregnancy occurrence rate with PPV and NPV: 30.78%, report: 33.02%

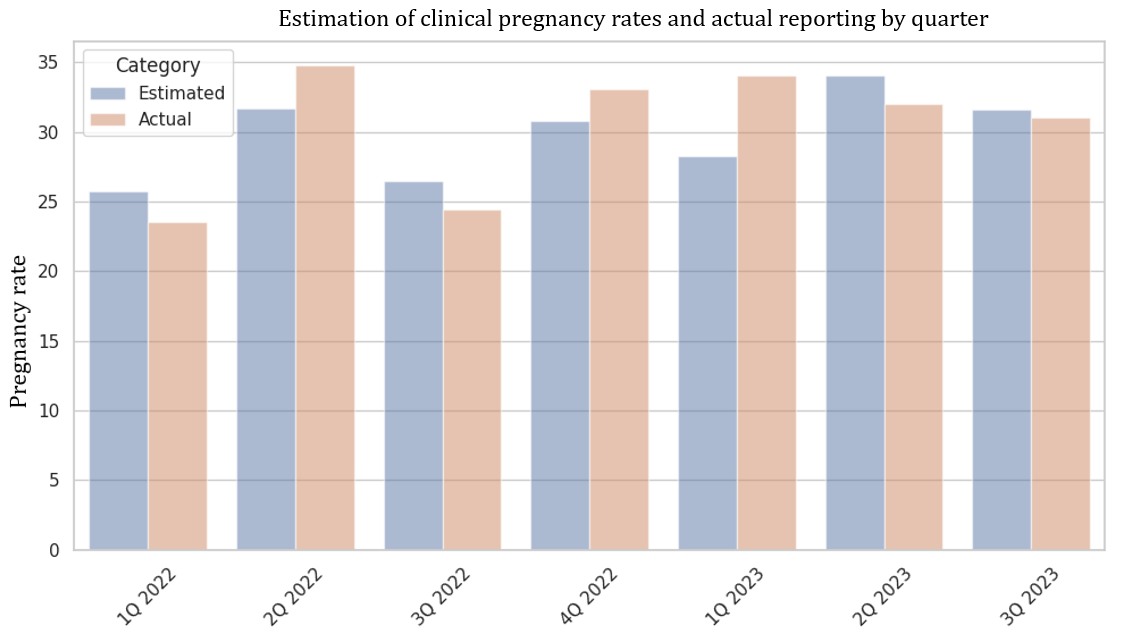
1st Quarter 2023: Estimated pregnancy occurrence rate with PPV and NPV: 28.21%, report: 34%

2nd Quarter 2023: Estimated pregnancy occurrence rate with PPV and NPV: 34.05%, report: 32%

3rd Quarter 2023: Estimated pregnancy occurrence rate with PPV and NPV: 31.56%, report: 31%

The average difference between the calculated predictions and reports was 2.56% that were not significant (U-Statistics = 20.0, p value = 0.62). These results convincingly demonstrate the effectiveness and reproducibility of the model and the prediction algorithm for the outcome of transfers embedded.

External validation of the model on the dataset from another clinic: 49662 embryos, 6240 protocols was used. Real PR 28.43%, Predicted PR 26.09% (p-value: 0.2185), ROC-AUC = 0.73, Test Accuracy: 76.34%.



**Areas of Application of the Model:**

The developed model allows for a detailed analysis of selected protocols with unsuccessful pregnancy outcomes while simultaneously providing a real-time working system for assessing the Key Performance Indicators (KPIs) of the laboratory. The model monitors statistical key performance indicators to evaluate the efficiency of individual embryologists and culture conditions. Discrepancies detected in the model's report can be used for systematic and early identification of unfavorable results and clinically significant changes in laboratory performance indicators.

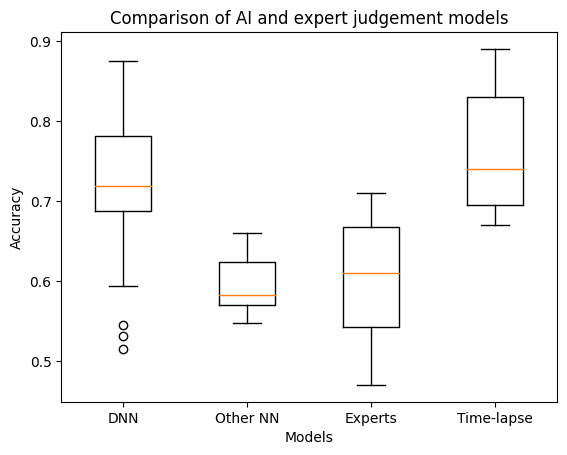
The Model AUC ( 0.68; SD = 0.054) implies that the neural network model not only excels in its predictive capabilities but also demonstrates efficiency in terms of processing speed when compared to alternative AI models discussed in the literature including commercial time-lapse systems (t-Statistics: -0.74, p-value = 0.46). The use of this neural network model is beneficial in terms of accuracy and interpretability, and it also stands out in terms of time efficiency, making it a favorable choice in practical applications.

The accuracy data of the model's outcome predictions surpass the analogous accuracy of predicting transfer outcomes by experts based on clinical and laboratory parameters—51% (range 43–59%) according to literature data. This enables its application for remote monitoring of the ART department's performance and serves as an effective mechanism for the clinic's overall quality control.

The developed deep learning neural network not only facilitates the necessary monitoring of internal quality control but also possesses sufficient predictive power for analyzing IVF protocols. It can be used for retrospective analysis of work results as well as for prospective evaluation of the chances of clinical pregnancy occurrence, proving to be promising in assisting medical decision-making.

For instance, the model performs predictions even for protocols where embryo transfer did not take place, but cryopreserved embryos are available. This model allows assessing the success chances of a cryo cycle in a particular program and deciding on the advisability of additional stimulation and a new IVF attempt or the prospect of cryotransfer, determining the need for further storage of vitrified embryos.

The model can be utilized as a tool for quality control and the success of ART programs. It provides insights into cases where a transfer should have resulted in pregnancy if the expected and actual pregnancy rates do not align—there are valid reasons for identifying factors hindering implantation. From the laboratory perspective, this involves monitoring KPIs, automatically calculated by the model; from the clinical side, it involves comparing the actual pregnancy rate with the theoretically calculated one.



**Conclusions:**

The developed deep learning neural network model successfully passed all validation tests, demonstrating comparable capabilities to AI systems used in ART for correctly predicting the chance of clinical pregnancy occurrence in a specific IVF protocol and the overall pregnancy rate over a selected period. The model allows for detailed tracking of laboratory KPIs and their impact on the transfer outcome.

**References:**

1. Kaufmann, S J et al. “The application of neural networks in predicting the outcome of in-vitro fertilization.” Human reproduction (Oxford, England) vol. 12,7 (1997): 1454-7. doi:10.1093/humrep/12.7.1454
2. Vogiatzi, Paraskevi et al. “An artificial neural network for the prediction of assisted reproduction outcome.” Journal of assisted reproduction and genetics vol. 36,7 (2019): 1441-1448. doi:10.1007/s10815-019-01498-7
3. Louis, Claudio Michael et al. “Review of computer vision application in in vitro fertilization: the application of deep learning-based computer vision technology in the world of IVF.” Journal of assisted reproduction and genetics vol. 38,7 (2021): 1627-1639. doi:10.1007/s10815-021-02123-2
4. Bormann, Charles L et al. “Performance of a deep learning based neural network in the selection of human blastocysts for implantation.” eLife vol. 9 e55301. 15 Sep. 2020, doi:10.7554/eLife.55301
5. Salih, M et al. “Embryo selection through artificial intelligence versus embryologists: a systematic review.” Human reproduction open vol. 2023,3 hoad031. 15 Aug. 2023, doi:10.1093/hropen/hoad031
6. Liu, Xiaoyan et al. “Construction of the machine learning-based live birth prediction models for the first in vitro fertilization pregnant women.” BMC pregnancy and childbirth vol. 23,1 476. 27 Jun. 2023, doi:10.1186/s12884-023-05775-3
7. Chamayou, S., Patrizio, P., Storaci, G., Tomaselli, V., Alecci, C., Ragolia, C., Crescenzo, C., & Guglielmino, A. (2013). The use of morphokinetic parameters to select all embryos with full capacity to implant. Journal of assisted reproduction and genetics, 30(5), 703–710. https://doi.org/10.1007/s10815-013-9992-2
8. Dal Canto, M., Bartolacci, A., Turchi, D., Pignataro, D., Lain, M., De Ponti, E., Brigante, C., Mignini Renzini, M., & Buratini, J. (2021). Faster fertilization and cleavage kinetics reflect competence to achieve a live birth after intracytoplasmic sperm injection, but this association fades with maternal age. Fertility and sterility, 115(3), 665–672. https://doi.org/10.1016/j.fertnstert.2020.06.023
9. Basile, N., Vime, P., Florensa, M., Aparicio Ruiz, B., García Velasco, J. A., Remohí, J., & Meseguer, M. (2015). The use of morphokinetics as a predictor of  implantation: a multicentric study to define and validate an algorithm for embryo selection. Human reproduction (Oxford, England), 30(2), 276–283. https://doi.org/10.1093/humrep/deu331
10. Franco, José G Jr et al. “Key performance indicators score (KPIs-score) based on clinical and laboratorial parameters can establish benchmarks for internal quality control in an ART program.” JBRA assisted reproduction vol. 21,2 61-66. 1 Jun. 2017, doi:10.5935/1518-0557.20170016
11. Goyal, A., Kuchana, M., & Ayyagari, K. P. R. (2020). Machine learning predicts live-birth occurrence before in-vitro fertilization treatment. Scientific reports, 10(1), 20925. https://doi.org/10.1038/s41598-020-76928-z
12. Raef, Behnaz et al. “Computational prediction of implantation outcome after embryo transfer.” Health informatics journal vol. 26,3 (2020): 1810-1826. doi:10.1177/1460458219892138
13. Blank, Celine et al. “Prediction of implantation after blastocyst transfer in in vitro fertilization: a machine-learning perspective.” Fertility and sterility vol. 111,2 (2019): 318-326. doi:10.1016/j.fertnstert.2018.10.030
14. Loewke, Kevin et al. “Characterization of an artificial intelligence model for ranking static images of blastocyst stage embryos.” Fertility and sterility vol. 117,3 (2022): 528-535. doi:10.1016/j.fertnstert.2021.11.022
15. Fréour, T., Le Fleuter, N., Lammers, J., Splingart, C., Reignier, A., & Barrière, P. (2015). External validation of a time-lapse prediction model. Fertility and sterility, 103(4), 917–922. https://doi.org/10.1016/j.fertnstert.2014.12.111
16. Berntsen, J., Rimestad, J., Lassen, J. T., Tran, D., & Kragh, M. F. (2022). Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences. PloS one, 17(2), e0262661. https://doi.org/10.1371/journal.pone.0262661
17. Reignier, A., Girard, J. M., Lammers, J., Chtourou, S., Lefebvre, T., Barriere, P., & Freour, T. (2019). Performance of Day 5 KIDScore™ morphokinetic prediction models of implantation and live birth after single blastocyst transfer. Journal of assisted reproduction and genetics, 36(11), 2279–2285. https://doi.org/10.1007/s10815-019-01567-x
18. Tran, D., Cooke, S., Illingworth, P. J., & Gardner, D. K. (2019). Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer. Human reproduction (Oxford, England), 34(6), 1011–1018. https://doi.org/10.1093/humrep/dez064
19. VerMilyea, M., Hall, J. M. M., Diakiw, S. M., Johnston, A., Nguyen, T., Perugini, D., Miller, A., Picou, A., Murphy, A. P., & Perugini, M. (2020). Development of an artificial intelligence-based assessment model for prediction of embryo viability using static images captured by optical light microscopy during IVF. Human reproduction (Oxford, England), 35(4), 770–784. https://doi.org/10.1093/humrep/deaa013
20. Bori, L., Meseguer, F., Valera, M. A., Galan, A., Remohi, J., & Meseguer, M. (2022). The higher the score, the better the clinical outcome: retrospective evaluation of automatic embryo grading as a support tool for embryo selection in IVF laboratories. Human reproduction (Oxford, England), 37(6), 1148–1160. https://doi.org/10.1093/humrep/deac066
21. Nelson, S. M., & Lawlor, D. A. (2011). Predicting live birth, preterm delivery, and low birth weight in infants born from in vitro fertilisation: a prospective study of 144,018 treatment cycles. PLoS medicine, 8(1), e1000386. https://doi.org/10.1371/journal.pmed.1000386