Time Series Classification for Modality-Converted Videos: A Case Study on Predicting Human Embryo Implantation from Time-Lapse Images

Mehryar Abbasi*†, Parvaneh Saeedi*‡, Jason Au§¶, and Jon Havelock§‖
*School Of Engineering Science,Simon Fraser University, Burnaby, BC, Canada
§Pacific Centre for Reproductive Medicine, Burnaby, BC, Canada
Email:†mabbasib@sfu.ca, ‡psaeedi@sfu.ca,¶jau@pacificfertility.ca,‖jhavelock@pacificfertility.ca

***Abstract*—Video analysis requires both spatial and temporal data analysis, unlike image analysis, which is limited to processing only spatial information. The added complexity for analyzing videos translates into the requirement for more sophisticated algorithms with diverse data to optimize a model. Creating large video datasets for analysis tasks is challenging, especially in the medical field, where data accessibility is limited. As a result, many computationally complex methods, such as 3D-CNN models, are not well-suited for medical applications. Therefore, innovative strategies are required to train deep learning (DL)-based models for limited video data. This paper proposes a system to predict human embryo implantation outcome in In Vitro Fertilization (IVF) process by analyzing image sequences captured during incubation. However, data availability is restricted for this task as acquiring and annotating data involves several complex steps. The proposed approach focuses on utilizing morphological changes of embryos over time and linking them to the outcome. We convert embryonic microscopic videos into multivariate time series arrays and apply state-of-the-art time series classifiers to predict growth patterns and outcomes. However, these classifiers fail to utilize the temporal patterns in the data and result in poor performance. Therefore, we propose to modify the time series classifiers with attention mechanisms that can capture both short- and long-term dependencies and improve the accuracy of predicting the success of the IVF procedure. The proposed method[1] demonstrates promising results, improving the prediction accuracy by 3.3% for Day 3 and 3.1% for Day 5 embryo time-lapse videos. Our ensemble classifier achieved a prediction accuracy of 77.5%, a 5.2% improvement over the state-of-the-art.**

***Index Terms*—component, formatting, style, styling, insert**

I. Introduction

In Vitro Fertilization (IVF) is the most common treatment for infertility [1]. The success rate of IVF treatment is approximately 35% [1]. In IVF, a patient's ovaries are hyper-stimulated with hormone injections followed by retrieval of multiple ova (human eggs). Fertilized ova (embryos) are cultured inside temperature-controlled incubators, where they are monitored and imaged at fixed time intervals. Embryos are ranked at specific times according to their morphological attributes and development progress. After 3 to 5 days, the highest quality embryo(s), as identified by expert embryologists, is transferred into the patient's uterus or frozen for later. Day 3 embryo assessment is performed according to two criteria: (1) number of cells, (2) quality of the cells [2]. In the Day-5 assessment, the quality of an embryo is assessed using characteristics of its three main structures: Trophectoderm's (TE), Zona Pellucida (ZP), and Inner Cell Mass (ICM) [3]. The grading scheme for Day-5 embryos has three scores: (1) the expansion level of the embryonic cavity, (2) the distinctiveness and compactness of ICMs cells, and (3) TE cell regularity [3]. Figure 1 (a) and (b) display a Day-3 and a Day-5 embryo, respectively.

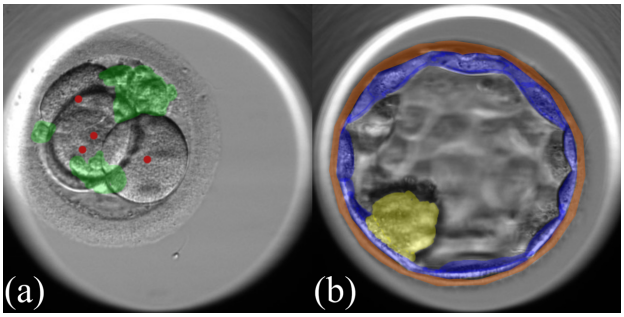

Fig. 1. (a) A day-3 human embryo (red indicates cell centroids and green fragmentation). (b) A day-5 human embryo (yellow, blue, and orange indicate TE, ICM, and ZP).

Embryo grading is performed by an expert embryologist, making it subjective, time-consuming, and expensive. Integrating artificial intelligence (AI) into IVF and embryo assessment has significantly improved the accuracy and efficiency of these processes. By leveraging AI, medical practitioners can predict which embryos have the highest probability of successful implantation and subsequent pregnancy, thereby minimizing the need for multiple rounds of IVF and reducing the financial and emotional strain on patients.

Embryologists currently use a method of embryo grading that involves evaluating embryos at a fixed time each day and inferring their overall quality throughout their incubation period through quality interpolation based on daily observa-

[1]https://github.com/mehryar72/Embryo-TSC

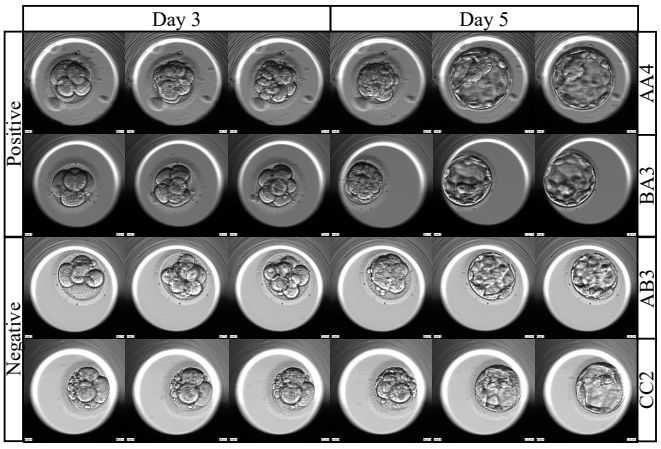

Fig. 2. Each row presents Day-3 and Day-5 samples of an embryo image sequence along with its Day 5 grade and implantation outcome.

tions. Such an approach could and would result in the loss of potentially valuable qualitative information, such as the onset of certain embryo development stages correlated with implantation likelihood [4] and over-time developmental patterns. Utilizing AI for embryo video processing is an innovative solution to tackle the mentioned issue. However, this task involves complex methods requiring significant training data to accurately process both temporal and spatial information. Obtaining videos is a costly and challenging process, and annotating them is time-consuming, leading to limited data availability. A viable approach could be to separate the spatial and temporal processing and training tasks. A model is trained to extract spatial features and convert the videos into a new modality first, and a second stage processes temporal information. However, commonly used methods for this step have yet to overtake the existing spatial-only analysis models.

In response to the challenges presented, this paper proposes a novel approach that leverages Time Series Classifiers (TSC) to utilize informative patterns over different time scales. Our proposed method trains two deep convolutional neural networks (CNNs) that act as spatial feature extractors. One model is trained on Day 3 images, while the other is on Day-5. Our previous research has demonstrated that separating the spatial analysis of different embryo stages can lead to improved prediction accuracy due to the unique and substantially different morphological attributes of each stage [5], [6]. By applying these trained models, we extract spatial features from each frame and transform the data modality from time-lapse images to multivariate time series arrays. We examine the application of TSC to correlate long and short-range embryonic developmental patterns with implantation outcomes.Initially, the state-of-the-art deep learning-based TSC performed poorly, despite being the best high-performing models [7], [8] on other time series classification tasks [9], [10]. Therefore, we had to modify these networks with a new self-attention block. In the new attention block, the content-based global self-attention is combined with a custom Gaussian-based localized attention to capture and combine the short-range and long-range temporal dependencies of our dataset.The results reveal a notable increase in the accuracy of pregnancy prediction. We also propose an ensemble classifier system that leverages Day-3 and Day-5 time series analysis results and Day-5 spatial processing to enhance the prediction accuracy even further.

II. Related Works

A. Deep Learning in Embryo Quality Assessment

Earlier studies on DL-based embryo quality assessment methods primarily aimed to develop models for automatic embryo grading [11]–[13]. One study by [13] aimed to extend grading to blastocyst image sequences using a CNN+RNN model. Unfortunately, the inclusion of a temporal analysis model resulted in a lower accuracy compared to the single-image grading method proposed in [12].

Recent research has focused on directly predicting outcomes (implantation or pregnancy test results [14]–[16] or live-birth results [17]–[21]), eliminating the need for embryo grading. While [14], [15], [17] used single blastocyst image analysis, [16], [18]–[20] used embryo time-lapse sequences. However, all [16], [18]–[20] suffer from common issues related to using highly imbalanced datasets and only recording the Area Under Curve (AUC), which may not provide an accurate performance measurement when working with heavily imbalanced data.

[18] achieved an AUC of 0.93 on a dataset of 8142 negative and 694 positive samples. [21] reported a 0.96 AUC on a highly unbalanced dataset of 15434 embryos, from which 12405 samples were discarded embryos labeled as negative samples. [16] used transfer learning to convert a pre-trained CNN+LSTM model to an embryo implantation predictor on a 216 negative and 56 positive samples dataset, achieving an AUC of 0.82. [19] trained a Resnet56 [22] to produce a live birth confidence score on a 379 negative and 91 positive samples dataset, resulting in an AUC of 0.642. However, their confidence score interquartile range indicated a failure to distinguish positive and negative samples. Finally, [20] used an I3D [23] model to predict the live-birth outcome on a dataset of 2212 videos (plus 15037 discarded embryos) and reported a separate AUC of 0.68. Their dataset's class imbalance included a ratio of 70% negative to 30% positive samples.

Insufficient data and overfitting pose significant challenges in previous studies, even in the case of 2D deep learning models that had to solely analyze spatial information for feature extraction or classification. Due to these issues, the utilized models were constrained to basic, shallow CNN architectures, as the more advanced and intricate models proved to be unfeasible. This limitation is clearly observed in the comparison between [6] and [5]. showing that employing fewer residual blocks in the models led to improved accuracy, highlighting the occurrence of overfitting as the complexity of the model increased.

Despite attempts to utilize deep learning for processing both temporal and spatial information in embryo videos, previous approaches have proven unsuitable for predicting outcomes.

RNN-based models are not a popular or a good choice for time series classification [9]. On the other hand, 3D convolutional models, such as I3D [23], require a large amount of training data and cannot model long-term temporal dependencies [24].

Here, we utilized the three most popular DL-based TSC models, FCN [7], ResNet [7], and InceptionTime [8].

B. DL-based Time Series Classifers

The field of time series classification has traditionally relied on non-deep learning (DL) methods, such as [25]. However, in recent years, the widespread availability of GPU-accelerated computation and the ease of access to such hardware have led to a surge in the popularity of DL-based approaches, mirroring trends in other domains. A notable limitation of traditional TSC methods is their inability to leverage GPUs for training, which severely hampers scalability [8], [25]. In contrast, DL-based TSC methods can effectively utilize available hardware resources, accelerating both training and inference processes, and thereby enhancing classifier scalability. Early DL-based TSC approaches employed relatively simple architectures, including Multi-Layer Perceptron (MLP), shallow 1D-CNNs, and Recurrent DNNs [9]. However, these networks exhibited inferior performance compared to traditional methods on widely used time series classification benchmarks, such as the UCR benchmark consisting of 128 problems [26]. The advent of more complex 1D-CNN architectures, such as Fully Convolutional Networks (FCN) and Residual Networks (ResNet), demonstrated that DL methods could achieve comparable results to traditional approaches with lower computational complexities and shorter training times. FCN, a three-layered 1D-CNN model, and ResNet, a deeper model with eleven 1D convolutional layers organized into three residual blocks, maintained the input series' length throughout their layers. InceptionTime emerged as a DL-based TSC method that achieved accuracy comparable to the best traditional techniques. It leveraged multiple inception modules [27], which simultaneously applied four convolutional filters of different kernel sizes to the input. FCN, ResNet, and InceptionTime have consistently ranked among the most successful TSC methods, achieving top positions [9], [28] on the UCR archive [26]. Consequently, these models were chosen as the baseline for our work. However, in their original forms, these models hindered the overall architecture's performance, necessitating modifications to enhance their effectiveness.

III. Methodology

The architecture of the proposed DL-based model is shown in Fig. 3. It comprises three stages: image feature extraction, time-series analysis, and ensemble voting. Its input is a series of Day-3 and Day-5 frames from an embryo sequence to predict the implantation outcome for that embryo.

A. Image feature extraction

We trained two separate CNN structures, D3FS and D5FS, to extract feature arrays from the captured Day-3 and Day-5 input frames. The combination of D3FS/D5FS with a fully connected (FC) layer was trained to predict implementation outcomes.

B. Time-series analysis

Temporal analysis of the input data is performed in the second stage of our embryo analysis architecture (orange block of Fig. 3). This stage takes two main inputs: multivariate time series feature arrays created by stacking the frame feature arrays generated from D3FS/D5FS outputs. A third univariate input is created by stacking the Day-5 binary prediction outcome across the temporal dimension. This input is not subject to training and is included solely to ensure an odd number of inputs for the final stage's majority voting. The selection of Day-5 binary prediction over Day-3 binary prediction was based on its stronger correlation with the implantation outcome, as supported by studies such as [5], [29].

We comprehensively assessed the effectiveness of multiple cutting-edge DL-based TSC models. Our evaluation included FCN, ResNet (RES) [7], InceptionTime (IT64) [8] and their modified variants. These modifications were applied by adding a self-attention encoder [30] with/without Positional Encoding (PE) before the Global Average Pooling (GAP) layer (Fig. 4).

In the Modified Encoder (ME) formulation, the attention matrix A was modified by combining it with A_2 using Eq. (1). $A_2 \in \mathbb{R}_+^{N\times N}$ encodes temporal placement information through an asymmetrical Gaussian distribution, where each value in row i and column j corresponds to distance-based attention generated using Eq. (2) between the i_{th} and j_{th} time samples. $F \in \mathbb{R}^{N\times d}$ represents the output time series array obtained from either the FCN, ResNet, or IT64 models. $K, Q, V \in \mathbb{R}^{N\times d}$ represent the Key, Query, and Value. $W_K, W_Q, W_V \in \mathbb{R}^{d\times d}$ and $W \in \mathbb{R}^{1\times d}$ are the learn-able weight matrices. $d,\ N$ represents the input dimension size and the sequence length, respectively. $\mathbb{S}$ scales each matrix row to have a maximum value of 1, while $\mathbb{N}$ normalizes each matrix row by dividing it by the sum of all its elements.

$$
\begin{aligned}
&K^T = W_K F^T,\ Q^T = W_Q F^T,\ V^T = W_V F^T,\\
&A = Softmax\left(\frac{QK^T}{\sqrt{d}}\right),\\
&A_m = \mathbb{N}\left(\frac{\mathbb{S}(A) + A_2}{2}\right), and\\
&O = A_m V.
\end{aligned}
\tag{1}
$$

In Eq. (2), P_i denotes the i_{th} row of A_2, and σ_i^2 represent the variance. b is a bias of 0.5, determined through external experimentation.

$$
\begin{aligned}
\sigma_i &= |W v_i| + b,\\
p_{i,j} &= e^{-\frac{1}{2}\frac{(i-j)^2}{\sigma_i^2}},\\
P_i &= (p_{i,1},\ p_{i,2}, p_{i,3},\ ...,\ p_{i,N}),\\
A_2^T &= (P_1, P_2, \ldots, P_T).
\end{aligned}
\tag{2}
$$

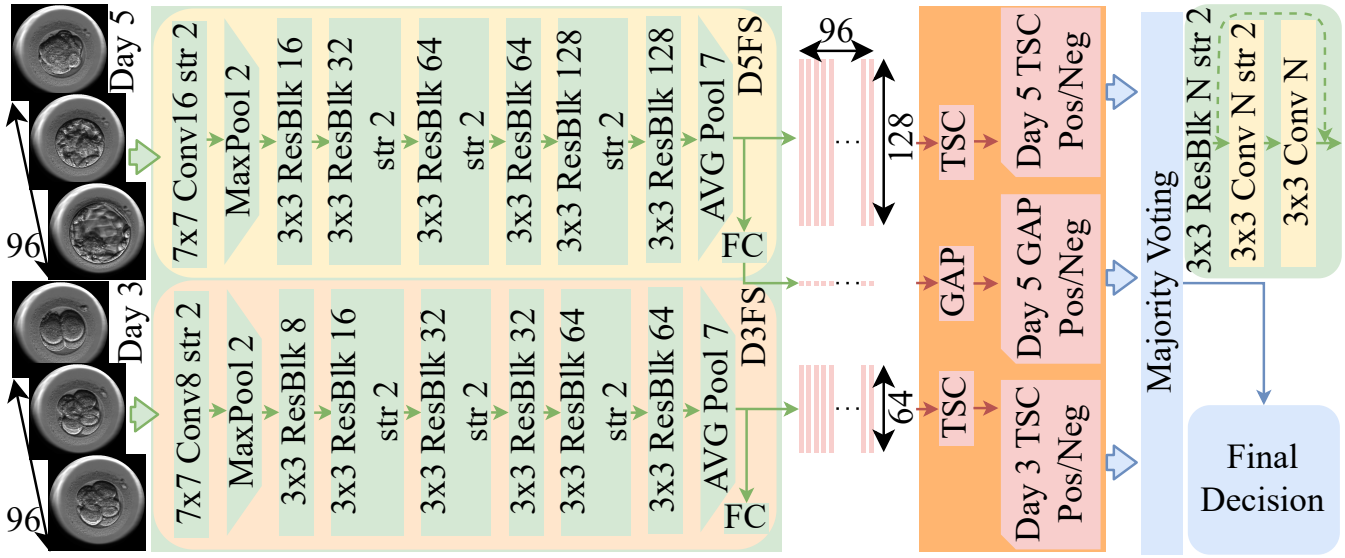

Fig. 3. Proposed 3-stage embryo analysis model. Green) Image feature extraction, Red) Time series analysis, Blue) Ensemble voting.

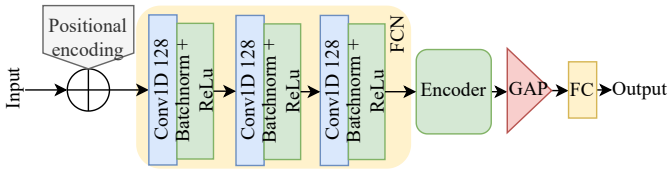

Fig. 4. Diagram of a self-attention augmented TSC using FCN.

The matrix A_2 effectively captures localized short-range patterns by utilizing a Gaussian distribution centered around each sample. Previous research, such as [31], has demonstrated that the dispersion of temporal information across the time dimension manifests as key segments with Gaussian-distributed characteristics. The P distributions are influenced by two factors: the content value and the distance between the samples. The content value affects the distribution variance at each time sample through a learnable weight denoted as W. Additionally, the relative positional placement information between samples contributes to the shape of the Gaussian distribution. By considering these dependencies, A_2 can effectively capture and represent both the localized short-range patterns and the relative positioning information present in the time series data. Meanwhile, A captures long-range dependencies through content-based self-attention. The scaling factor $\mathbb{S}$ is applied to ensure that the magnitudes of A and A_2 are within the same range and their impacts are balanced. Subsequently, the normalization operation $\mathbb{N}$ is employed to adjust the attention values, ensuring that they are appropriately scaled and distributed. This normalization step is crucial for maintaining the distribution properties of the attention mechanism.

C. Ensemble Voter

In the final stage, The information extracted from Day-3 and Day-5 sequences are combined using majority voting to generate an outcome prediction. The ensemble voter takes in three inputs: the classification outputs for the Day-3, Day-5 time series, and the Day-5 average frame score. The final output is determined by selecting the prediction with the highest number of votes. We utilized a majority vote approach rather than aggression for several reasons. Firstly, each TSC path produced outputs with varying dynamic ranges and scales due to their separate training. Secondly, we aimed to avoid the need for additional training by introducing a new learnable weighted aggregator method. Thirdly, we had an existing, freely available classification output that could be incorporated into the final stage without incurring any extra costs. Furthermore, this additional output (Day-5 average frame score) demonstrated a strong correlation with the desired outcome.

IV. Experiments

A. Data

Our human embryo dataset comprises 130 time-lapse image sequences collected using an $EmbryoScope^{TM}$ time-lapse incubator (Vitrolife) with a 15-minute acquisition interval (2 presents a few samples of this dataset). There are 60 positive implantation outcome sequences (46%) and 70 negative implantation outcome sequences (54%). We classify and automatically extract images captured between hours 48-72 post-fertilization as "Day-3 sequences," while those captured after hour 96 are labeled as "Day-5 sequences." The Day-3 sequences consist of a fixed 96 frames, whereas the Day-5 sequences vary in length, ranging between 67 and 96 frames. To standardize the input for the Day-5 TSC model, the sequences were zero-padded to create 96-frame sequences.

The Day-5 sequences were truncated due to the transfer of day 5 embryos to the patients based on the discretion of the embryologist.The dataset contains 12,480 images for Day-3 sequences and 10,097 for Day-5 sequences. The image format is 8-bit with 500×500 pixels. Images are automatically translated, Such that the embryo is centered in each frame [6], and the background region is set to zero. Images are downsampled to 224×224 pixels. The dataset is divided into five groups of 26 sequences. We used 5-fold cross-validation for the training and testing of our models.

B. Setup Configurations and Hyper Parameter

The first step is to train the spatial feature extractor models, which act as modality conversion models. The D3SF and D5SF were trained independently, with one Day-3 and one Day-5 model trained for each of the 5 folds, resulting in 10 models (5 pairs). Frames were assigned a ground truth label of positive or negative based on the actual pregnancy outcome, and binary cross-entropy loss was used to train the models. The training was done over 100 epochs using a learning rate of $1e{-}4$ and an Adam optimizer, as in [5]. Additionally, during the training process, some image augmentations were applied, such as random rotation (between 0-45 degrees), random image flipping (up/down), and random downscaling (ratio between 0.85 - 1). The batch size was limited to the number of training sequences, which was 104.

The TSC models were optimized using binary cross-entropy loss and Adam optimizer with a learning rate of $1e{-}6$ for 1000 epochs with a batch size of 16. Each TSC model was trained and tested five times for each of the five data folds, resulting in 25 instances of each model. The reported results in the following section are the average performance over 25 instances.

We also tested a bi-directional LSTM model on the Day-3 and Day-5 multivariate TSC arrays as part of our experiments. The hidden size of the bi-directional LSTM was set to match the dimension of the input array, 64 for Day-3 and 128 for Day-5 (Sequence length is the same).

We tested the I3D model, which does not require a feature extraction stage. The I3D model only captures temporal dependencies in the window of 32 frames. We had to randomly select 32 sequential frames from each training video for each epoch. During the inference stage, I3D takes in the whole image and it aggregates the predictions. To work within the limitations of GPU memory, we set the batch size to 16.

The majority voter model, which does not require further training, relies on TSC models for Day-3 and Day-5. We selected the best-performing TSC model for each day to set up the majority voter model. The crucial point is that we trained 5 instances of a TSC model for each data fold. To ensure comprehensive testing of the majority voter, we repeated the testing operation 25 times for each data fold, once for each combination of Day-3 and Day-5 instances. This resulted in a total of 125 experiments. The reported results for the majority voter model are the average of all 125 experiments.

All experiments are performed on a Compute Canada [32] node with an NVIDIA V100 Volta GPU (32G) unit.

C. Results

TABLE I
HUMAN EMBRYO OUTCOME PREDICTION ON DAY-3 AND DAY-5 IMAGE SEQUENCES.

	Day-3				Day-5			
Model	PER [1]	RCL [2]	JI [3]	ACC [4]	PER [1]	RCL [2]	JI [3]	ACC [4]
FC+GAP [5]	73.0	70.7	53.5	70.0	74.4	71.8	55.9	72.3
I3D	64.9	61.7	41.0	61.8	68.3	64.4	38.8	65.4
Bi-LSTM	68.2	65.6	48.3	65.8	27.9	49.8	36.0	51.8
FCN	68.6	68.4	50.9	68.2	72.4	71.7	55.9	71.8
FCN+ME+PE	73.9	72.7	57.2	72.9	75.3	74.5	59.4	74.6
FCN+ME	**76.9**	**73.9**	**58.3**	**74.2**	75.3	73.4	57.6	73.5
RES	68.3	66.5	48.6	66.8	74.3	72.9	56.7	72.6
RES+ME+PE	73.0	70.6	54.3	70.8	74.4	72.2	56.4	72.6
RES+ME	72.4	70.6	54.2	70.5	75.1	72.8	57.0	73.1
IT64	61.1	63.7	44.6	63.7	65.7	67.9	49.5	67.3
IT64+ME+PE	71.8	70.4	54.0	71.0	**76.9**	**75.1**	**59.9**	**75.2**
IT64+ME	69.3	69.7	53.3	70.0	74.7	73.3	57.9	73.7

[1] Precision [2] Recall [3] Jaccard-index [4] Accuracy

Table I displays the results of experimenting with different classifiers on sequences from each day. The first row of the table shows the results obtained by averaging the output of the spatial classifier (D3SF/D5FS + FC) over time. This is achieved by applying a Temporal Global Average Pooling (GAP) layer to the output of the Spatial Classifier. These results demonstrate that the performance of the base TSC models, without attention augmentation, is inferior to that of the no temporal analysis model of each day (i.e., "FC+GAP" and "FC+GAP").

The "I3D" and "LSTM" models, utilized in previous studies, demonstrated the lowest performance among all the methods examined. Surprisingly, both models underperformed even when compared to the base method, which lacks any form of temporal processing. Furthermore, they also exhibited inferior performance compared to all other evaluated temporal processing methods.

The use of the TSC classifiers "FCN+ME" and "IT64+ME+PE" resulted in an increase in the classification accuracy by 4.2% and 2.9% for Day-3 and Day-5 sequences, respectively, compared to the previous state-of-the-art method [5], which lacked temporal analysis. As a result, these TSC classifiers were selected for the final model on their respective days.

The results of our Majority Voter model are shown in Table II. This proposed model includes image feature extractors, time series analysis, and ensemble voting. It is 6.6% more accurate than the "Blastocyst prediction" model from [14], which uses single blastocyst stage images as input. The proposed

TABLE II
EMBRYO IMPLANTATION OUTCOME PREDICTION RESULTS.

Prediction	Method	PER [1]	RCL [2]	JI [3]	ACC [4]
Live-birth	CNN+ CEE [17]	70.2	71.4	55.3	74.3
Implantation	HFC [15]	61.5	60.5	44	62
	Multi-stream CNN [14]	71.1	72.7	56	70.9
	Day-3 GAP [5]	73.0	70.7	53.5	70.0
	Day-5 GAP [5]	74.4	71.8	55.9	72.3
	Day-5 Blastocyst GAP [6]	74.1	72.6	57.2	73.8
	Day-3 TSC	76.9	73.9	58.3	74.2
	Day-5 TSC	76.9	75.1	59.9	75.2
	Majority Vote	**78.8**	**77.3**	**62.9**	**77.5**

[1] Precision [2] Recall [3] Jaccard-index [4] Accuracy

model is also 7.5% and 5.2% more accurate than individual "Day 3 FC+GAP" and "Day 5 FC+GAP" [5] models.

V. CONCLUSION

This paper presented the challenges of analyzing video data with data scarcity and complex long-range temporal patterns. It then conducted experiments to predict the pregnancy outcome of human embryo implantation in IVF using this type of data. We proposed a novel approach that converted video data into multivariate time series and utilized intelligent attention-augmented time series classifiers for outcome prediction. The proposed approach demonstrated a significant improvement in accuracy compared to the existing methods by delivering a prediction accuracy of 77.5% by combining Day-3 and Day-5 models. As a result, our approach is reliable for predicting IVF outcomes, aiding clinicians in making informed decisions, and improving success rates. In addition, the study highlighted the importance of customized solutions that address the unique challenges and limitations associated with specific data.

REFERENCES

[1] M. C. Inhorn and P. Patrizio, "Infertility around the globe: new thinking on gender, reproductive technologies and global movements in the 21st century," *Human reproduction update*, vol. 21, no. 4, pp. 411–426, 2015.

[2] K. Lundin, C. Bergh, and T. Hardarson, "Early embryo cleavage is a strong indicator of embryo quality in human IVF," *Human reproduction*, vol. 16, no. 12, pp. 2652–2657, 2001.

[3] C. Manna, L. Nanni, A. Lumini, and S. Pappalardo, "Artificial intelligence techniques for embryo and oocyte classification," *Reproductive biomedicine online*, vol. 26, no. 1, pp. 42–49, 2013.

[4] L. Rienzi, D. Cimadomo, A. Delgado, M. G. Minasi, G. Fabozzi, R. Del Gallego, M. Stoppa, J. Bellver, A. Giancani, M. Esbert *et al.*, "Time of morulation and trophectoderm quality are predictors of a live birth after euploid blastocyst transfer: a multicenter study," *Fertility and sterility*, vol. 112, no. 6, pp. 1080–1093, 2019.

[5] M. Abbasi, P. Saeedi, J. Au, and J. Havelock, "A deep learning approach for prediction of ivf implantation outcome from day 3 and day 5 time-lapse human embryo image sequences," in *2021 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2021, pp. 289–293.

[6] ——, "Timed data incrementation: A data regularization method for ivf implantation outcome prediction from length variant time-lapse image sequences," in *2021 IEEE 23rd International Workshop on Multimedia Signal Processing (MMSP)*. IEEE, 2021, pp. 1–5.

[7] Z. Wang, W. Yan, and T. Oates, "Time series classification from scratch with deep neural networks: A strong baseline," in *Int. joint Conf. Neural Netw.* IEEE, 2017, pp. 1578–1585.

[8] H. I. Fawaz, B. Lucas, G. Forestier, C. Pelletier, D. F. Schmidt, J. Weber, G. I. Webb, L. Idoumghar, P.-A. Muller, and F. Petitjean, "Inceptiontime: Finding AlexNet for time series classification," *Data Min. Knowl. Discov.*, vol. 34, no. 6, pp. 1936–1962, 2020.

[9] H. I. Fawaz, G. Forestier, J. Weber, L. Idoumghar, and P.-A. Muller, "Deep learning for time series classification: a review," *Data Min. Knowl. Discov.*, vol. 33, no. 4, pp. 917–963, 2019.

[10] M. Abbasi and P. Saeedi, "Enhancing multivariate time series classifiers through self-attention and relative positioning infusion," *arXiv preprint arXiv:2302.06683*, 2023.

[11] P. Khosravi, E. Kazemi, Q. Zhan, J. E. Malmsten, M. Toschi, P. Zisimopoulos, A. Sigaras, S. Lavery, L. A. Cooper, C. Hickman *et al.*, "Deep learning enables robust assessment and selection of human blastocysts after in vitro fertilization," *NPJ Digit. Med.*, vol. 2, no. 1, pp. 1–9, 2019.

[12] T.-J. Chen, W.-L. Zheng, C.-H. Liu, I. Huang, H.-H. Lai, and M. Liu, "Using deep learning with large dataset of microscope images to develop an automated embryo grading system," *Fertility & Reproduction*, vol. 1, no. 01, pp. 51–56, 2019.

[13] M. F. Kragh, J. Rimestad, J. Berntsen, and H. Karstoft, "Automatic grading of human blastocysts from time-lapse imaging," *Computers in biology and medicine*, vol. 115, p. 103494, 2019.

[14] R. M. Rad, P. Saeedi, J. Au, and J. Havelock, "Predicting human embryos' implantation outcome from a single blastocyst image," in *Int. Conf. of the IEEE Eng. in Med. and Biol. Society*. IEEE, 2019, pp. 920–924.

[15] A. Chavez-Badiola, A. F.-S. Farias, G. Mendizabal-Ruiz, R. Garcia-Sanchez, A. J. Drakeley, and J. P. Garcia-Sandoval, "predicting pregnancy test results after embryo transfer by image feature extraction and analysis using machine learning," *Scientific Reports*, vol. 10, no. 1, pp. 1–6, 2020.

[16] D. H. Silver, M. Feder, Y. Gold-Zamir, A. L. Polsky, S. Rosentraub, E. Shachor, A. Weinberger, P. Mazur, V. D. Zukin, and A. M. Bronstein, "Data-driven prediction of embryo implantation probability using ivf time-lapse imaging," *arXiv preprint arXiv:2006.01035*, 2020.

[17] Y. Miyagi, T. Habara, R. Hirata, and N. Hayashi, "Predicting a live birth by artificial intelligence incorporating both the blastocyst image and conventional embryo evaluation parameters," *Artif. Intell. in Med. Imag.*, vol. 1, no. 3, pp. 94–107, 2020.

[18] D. Tran, S. Cooke, P. Illingworth, and D. Gardner, "Deep learning as a predictive tool for fetal heart pregnancy following time-lapse incubation and blastocyst transfer," *Human Reproduction*, vol. 34, no. 6, pp. 1011–1018, 2019.

[19] Y. Sawada, T. Sato, M. Nagaya, C. Saito, H. Yoshihara, C. Banno, Y. Matsumoto, Y. Matsuda, K. Yoshikai, T. Sawada *et al.*, "Evaluation of artificial intelligence using time-lapse images of ivf embryos to predict live birth," *Reproductive BioMedicine Online*, vol. 43, no. 5, pp. 843–852, 2021.

[20] J. Berntsen, J. Rimestad, J. T. Lassen, D. Tran, and M. F. Kragh, "Robust and generalizable embryo selection based on artificial intelligence and time-lapse image sequences," *Plos one*, vol. 17, no. 2, p. e0262661, 2022.

[21] B. Huang, S. Zheng, B. Ma, Y. Yang, S. Zhang, and L. Jin, "Using deep learning to predict the outcome of live birth from more than 10,000 embryo data," *BMC Pregnancy and Childbirth*, vol. 22, no. 1, pp. 1–7, 2022.

[22] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *IEEE Conf. Comput. vision and pattern Recognit.*, 2016, pp. 770–778.

[23] J. Carreira and A. Zisserman, "Quo vadis, action recognition? a new model and the kinetics dataset," in *IEEE Conf. Comput. Vis. and Pattern Recognit.*, 2017, pp. 6299–6308.

[24] S. Arif and J. Wang, "Bidirectional lstm with saliency-aware 3d-cnn features for human action recognition," *Journal of Engineering Research*, vol. 9, no. 3A, 2021.

[25] M. Middlehurst, J. Large, M. Flynn, J. Lines, A. Bostrom, and A. Bagnall, "HIVE-COTE 2.0: a new meta ensemble for time series classification," *Mach. Learn.*, vol. 110, no. 11, pp. 3211–3243, 2021.

[26] H. A. Dau, A. Bagnall, K. Kamgar, C.-C. M. Yeh, Y. Zhu, S. Gharghabi, C. A. Ratanamahatana, and E. Keogh, "The UCR time series archive," *IEEE/CAA J. Autom. Sin*, vol. 6, no. 6, pp. 1293–1305, 2019.

[27] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *IEEE Conf. Comput. Vis. Pattern Recognit.*, 2015, pp. 1–9.
[28] K. Ouyang, Y. Hou, S. Zhou, and Y. Zhang, "Convolutional Neural Network with an Elastic Matching Mechanism for Time Series Classification," *Algorithms*, vol. 14, no. 7, p. 192, Jul. 2021.
[29] D. K. Gardner, M. Lane, J. Stevens, T. Schlenker, and W. B. Schoolcraft, "Blastocyst score affects implantation and pregnancy outcome: towards a single blastocyst transfer," *Fertility and sterility*, vol. 73, no. 6, pp. 1155–1158, 2000.
[30] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Adv. Neural Inf. Process. Syst.*, vol. 30, 2017, pp. 5998–6008.
[31] A. Piergiovanni and M. Ryoo, "Temporal gaussian mixture layer for videos," in *Int. Conf. Mach. Learn.*, 2019, pp. 5152–5161.
[32] https://www.computecanada.ca/.