International Journal of Medical Toxicology & Forensic Medicine
2026, Volume 16, E53285

Research Paper

Pep-Ensem-2n: A Robust Deep Ensemble Framework for Trimester-Wise Prediction of Preeclampsia Using Clinical Data

Sivaranjani Swaminathan[1,2*], **Aruna Rajendiran**[2]

1. *Department of Computer Science and Engineering, Pachaiyappa's College for Men, Kanchipuram, India.*
2. *Vel Tech Rangarajan Dr. Sagunthala R&D Institute of Science and Technology, Chennai, India.*

Citation Swaminathan S, Rajendiran A. Pep-Ensem-2n: A Robust Deep Ensemble Framework for Trimester-Wise Prediction of Preeclampsia Using Clinical Data. International Journal of Medical Toxicology and Forensic Medicine. 2026; 16:E53285.

https://doi.org/10.22037/ijmtfm.v16.53285

Article info:

Received: 27 August, 2026

Accepted: 24 September, 2026

Published: 01 November, 2026

Keywords:

Preeclampsia (PE), Machine Learning (ML), Deep Learning (DL), Ensemble Learning (EL), Clinical Decision Support System (CDSS), Random Forest (RF), Temporal Convolutional Network (TCN), Synthetic Minority Oversampling Technique (SMOTE), Medical Prediction

ABSTRACT

Background: There is significant maternal and perinatal morbidity associated with Preeclampsia. Maternal variability in physical and biochemical parameters during gestation makes accurate, timely risk assessment difficult. In this study, we propose a deep ensemble model, PEP-Ensem-2N, to predict preeclampsia at various pregnancy visits.

Methods: It incorporates missing-value imputation, standardization, random-forest feature selection, and class-balancing using the Synthetic Minority Oversampling Technique. We weight and soft-vote the prediction probabilities from five learning architectures (long short-term memory, gated recurrent unit, simple recurrent neural network, temporal convolutional network, and deep convolutional network). We used three datasets to evaluate the framework: PEP-T1, PEP-T2, and PEP-T3. We compared various model structures and activation functions and assessed the relative importance of the model inputs by comparing random forest feature rankings.

Results: The reported experiments showed that the rectified linear unit (ReLU)-based model PEP-Ensem-2N achieved an accuracy of 0.9946 on the PEP-T1 dataset, 0.9992 on PEP-T2, and 0.9990 on PEP-T3.

Conclusion: PEP-Ensem-2N is a feature-selection, class-imbalance, and probability-fusion approach within a unified prediction framework. The reported performance should be confirmed through counts and evaluation procedures. It needs independent validation with observed clinical data before it can be applied in clinical settings.

*** Corresponding Author:**
Sivaranjani Swaminathan, PhD
Department of Computer Science and Engineering, Pachaiyappa's College for Men, Kanchipuram, India.
E-mail: *sivaranjaniresearchscholar@gmail.com*

Copyright © 2026 The Author(s);
This is an open-access article distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0: https://creativecommons.org/licenses/by-nc/4.0/), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Introduction

Perinatal and maternal morbidity associated with hypertensive disorders of pregnancy (HDP), particularly preeclampsia (PE), is significant. The presentation and clinical characteristics are heterogeneous and evolving, making it problematic to determine risk in a timely manner. Machine learning is therefore an interesting approach for linking mothers' genetic characteristics with clinical and biochemical parameters. Previous studies have focused on intensive care requirements in women with PE [1], prediction in a middle-income country [2], or population-based risk prediction in China [3]. The case studies show the versatility of predictive modeling in obstetric applications. Researchers have also studied other sources of predictive information and clinical outcomes. Researchers have studied ECG signals for PE detection and prediction [4], and conducted prospective studies to predict PE and intrauterine growth restriction [5]. Other applications include predicting delivery within 7 days of early-onset PE diagnosis [6], incorporating maternal and fetal variables to evaluate PE and growth restriction [7], developing models using retrospective case–control data [8], and recognizing severe PE [9]. However, these methods cannot be readily compared directly because of differences in composition, prediction date, and the nature of the predictions. Biological and methodological research provides other directions for model development. Placental growth factor response to LMWH provides insight into biomarker behavior in pregnancies at risk of placental dysfunction [10]. Researchers have developed a deep neural network to address the imbalanced distribution of PE outcomes [11], and used placental ultrasound texture analysis to evaluate imaging features for predicting hypertensive disorders [12]. Machine learning is also used to assess cardiovascular risk after PE [13]. A similar obstetric study explored combining several predictors, using ensemble learning to predict postpartum hemorrhage in remote healthcare systems [14]. Another source for PE prediction via EHRs is information gathered during routine care [15]. In complementary biomarker research, researchers have used the maternal leptin/ceramide ratio for early risk assessment [16], while clinical applications of placental dysfunction have focused on angiogenic factors [17]. Prospective cohort modeling of PE and intrauterine growth restriction (IUGR) [18] also emphasizes the importance of matching the predictor to the desired outcome and gestational age, as well as matching the predictor to early pregnancy prediction [19] and models for late-onset PE [20]. Together, these studies support considering complementary models consistently during pregnancy visits. This evaluation should include transparent data provenance, participant-level partitioning, processing restricted to the training data, and assessment of discrimination, calibration, and uncertainty. In this study, we propose a new visit-specific ensemble method, PEP-Ensem-2N, that combines weighted averaging and soft voting based on Random Forest (RF) feature selection and Synthetic Minority Oversampling Technique (SMOTE). The architecture comprises hybrid models of long short-term memory, gated recurrent unit, simple recurrent neural network, temporal convolutional network, and deep convolutional network. The aim is to investigate the role of complementary learners, similar to PE risk modeling in pregnancy visits. Further validation with clinical data is required to validate the clinical applicability.

Materials and Methods

Figure 1 summarizes the end-to-end process of the proposed PEP-ENSEM-2N architecture, enabling accurate preeclampsia prediction and showing a pipeline from data acquisition to clinical decision-making. The workflow ensures rigorous data processing, trustworthy model training, and unbiased execution critique grounded in clinical understanding. In the proposed work, (https://github.com/PRAISE-Lab-Repository/PEPrML) employed (PEP-T1, PEP-T2, and PEP-T3); in turn, three publicly available synthetic demonstration datasets from PEPrML representing pregnancy visits were used for analysis. Each file contains 1000 records. The datasets have 56, 99, and 144 columns, respectively, including the identifier, PublicID, and the outcome variable (OUTCOME). After omitting these two columns, 54, 97, and 142 candidate predictors remain for further processing. These records still do not represent a clinical cohort of observed patients, so this analysis is performed on synthetic data. The datasets include structured clinical, biochemical, and demographic variables measured at various gestation periods, enabling longitudinal risk evaluation. PEP-T1 comprises 56 attributes, including primarily first-trimester biomarkers (ADAM12, PAPP-A, PLGF, and SFLT1), maternal demographic data, and early ultrasound results. PEP-T2 augments the feature space with second-trimester measurements, placental biomarkers, and some other clinical information, such as mean arterial pressure (MAPV2), and longitudinal biometric changes (LBCHG_V2).

The most comprehensive dataset, PEP-T3, has only 144 attributes and combines third-trimester

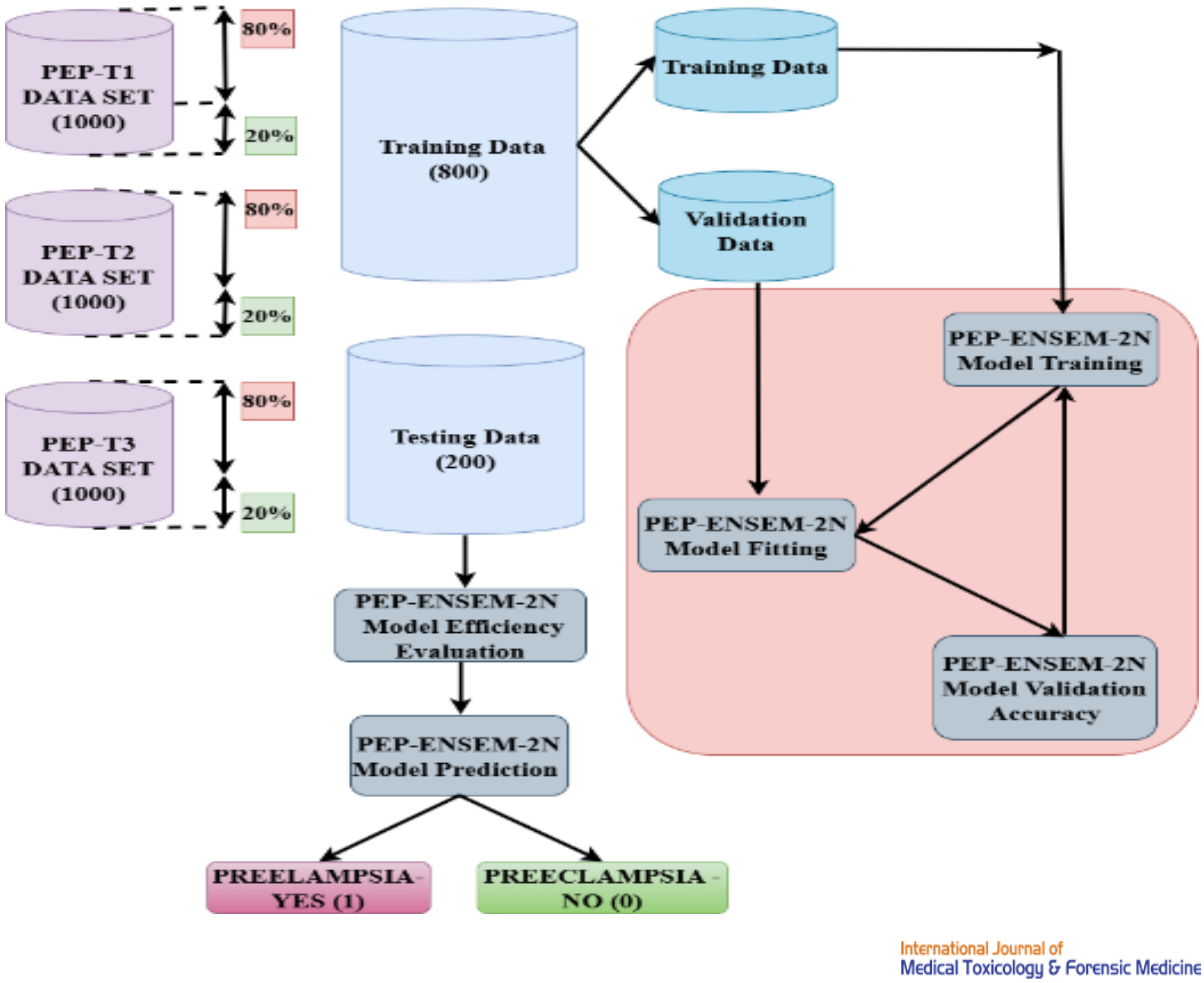


Figure 1. Schematic representation of the PEP-Ensem-2N prediction framework.

longitudinal biochemical patterns with cumulative risk features to represent late-pregnancy physiological dynamics. In all datasets, the patient's unique identifier is PublicID, and the outcome is a binary target variable (0 = no preeclampsia and 1 = Preeclampsia). We also add derived biomarker ratios, including SFLT1/PLGF, to increase predictive capacity and clinical relevance. To make sure that the performance evaluation is fair and unbiased, every dataset (PEP-T1, PEP-T2, and PEP-T3) is properly partitioned 80:20 stratified, where 80 percent of the observations are allocated to the model development phase, and 20 percent of the observations are reserved as an independent test set; this is required so that the results obtained are representative of how the model will behave in the real world on unseen data. The datasets are split into training, validation, and test sets to follow a consistent experimental protocol and ensure fair comparisons. We develop the model by training and refining the PEP-ENSEM-2N architecture and iteratively tuning its parameters. Validation accuracy and loss values provide feedback during weight tuning and the learning process. The optimized model is then tested on an independent subset of the data it was not trained on to derive unbiased estimates of predictive performance. Finally, the trained system is transferred to the prediction phase, whereby the patient records can be classified into two clinically interesting sets: Preeclampsia - YES (1) and Preeclampsia - NO (0). This binary output offers direct clinical interpretability and supports future risk stratification, helping healthcare professionals intervene early and plan individual treatment.

Figure 1 outlines the procedural workflow, while Figure 2 provides a detailed view of the internal architecture, giving an overall picture of the proposed PEP-ENSEM-2N framework. Figure 2 shows the detailed system architecture of the proposed PEP-ENSEM-2N model, focusing on the sequential processing stages that jointly support robust and precise preeclampsia prediction. The architecture begins with a preprocessing module that removes irrelevant attributes and duplicate records, reducing noise and redundancy. It handles missing values using statistical imputation methods to retain data completeness and encodes target labels numerically for supervised learning. The refined feature subset is then fed back to the model-building layer.

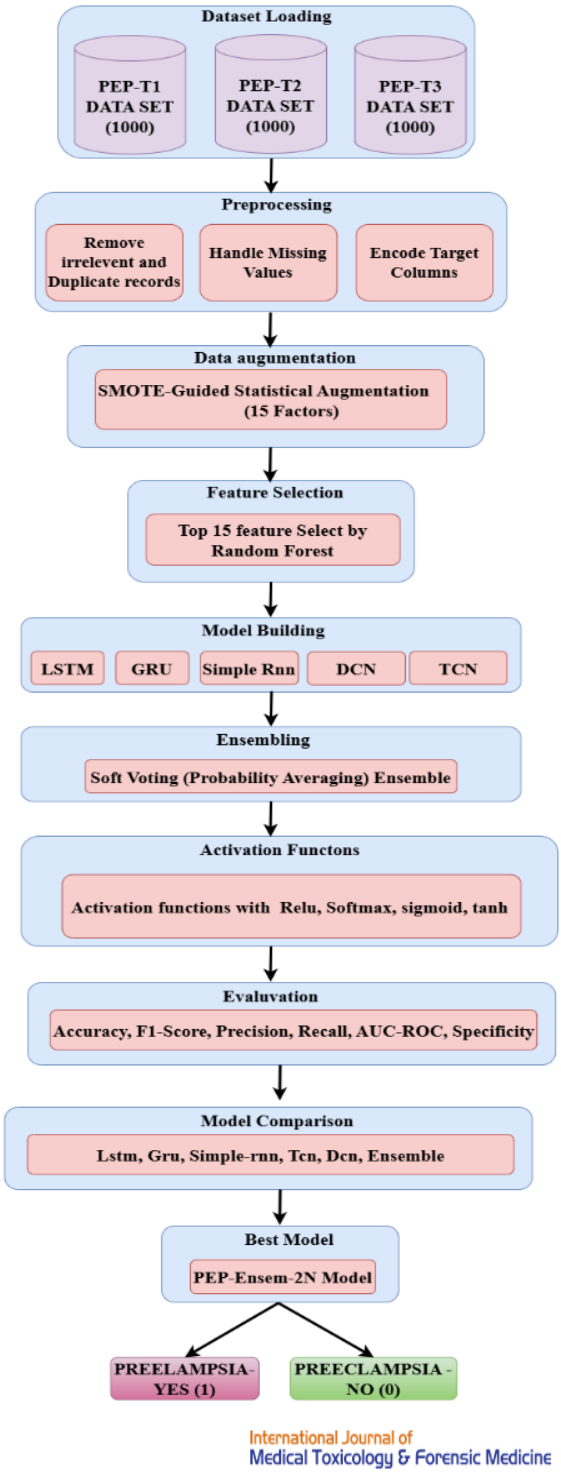
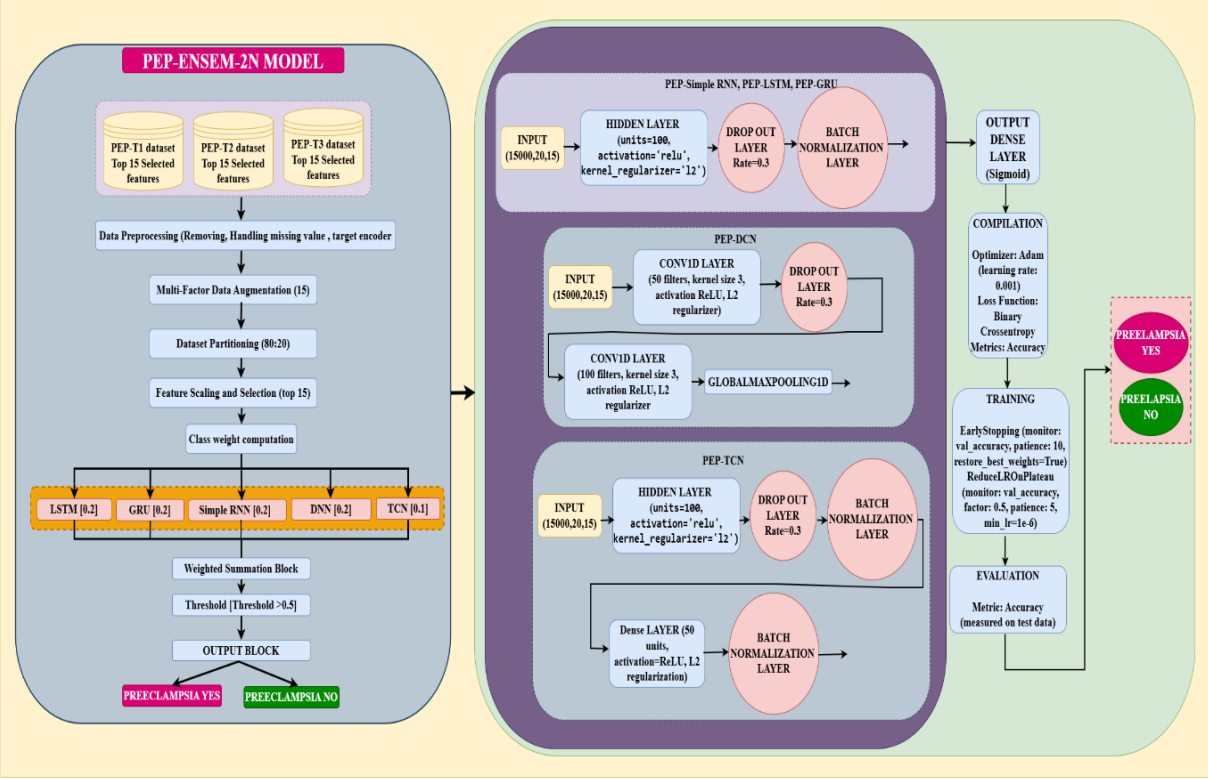


Figure 2. Overall architecture of the proposed PEP-Ensem-2N prediction framework.

The model-building layer contains a heterogeneous set of deep learning architectures, i.e., Long Short-Term Memory (LSTM), Gated Recurrent Unit (GRU), Simple Recurrent Neural Network (RNN), Deep Convolutional Network (DCN), and Temporal Convolutional Network (TCN). Each model learns complementary representations, capturing temporal dependencies, nonlinear relationships, and hierarchical feature abstractions. To make the most of these heterogeneous learners, we implement an ensemble learning module based on soft voting, that is, averaging the probabilities. The final prediction is obtained by a weighted soft-voting ensemble of all the base learners, explained in detail in the methodological framework of Section 3. This approach significantly boosts predictive accuracy and provides resilience to idiosyncratic model errors. Activation functions, such as ReLU, sigmoid, tanh, and softmax, are applied in both hidden and output layers to inject nonlinearity, control gradient flow, and optimize effectively. The terminal component of the architecture is an evaluation module that assesses model performance using a comprehensive set of metrics, including accuracy, precision, recall, F1-score, AUC-ROC, and specificity. These measures provide a well-rounded assessment of the diagnostic reliability, sensitivity for identifying cases with the disease, and overall clinical usefulness.

Implementation Setup of PEP-ENSEM-2N Model

Building on the methodological principles and architectural design described in the previous section, the PEP-ENSEM-2N framework is implemented as an end-to-end deep learning pipeline that systematically converts raw multi-trimester clinical datasets into robust, accurate, and clinically interpretable risk predictions for preeclampsia. The conceptual workflow is implemented through a reproducible computational architecture that includes data preprocessing, class imbalance mitigation, feature refinement, deep representation learning, and ensemble decision fusion. Figure 3 illustrates PEP-ENSEM-2N's overall implementation configuration, showing the progression from trimester-specific datasets to the final unified ensemble-based diagnostic decision through a series of processing steps. Each module is designed to ensure numerical stability, effective learning, and good generalization across heterogeneous patient populations. The implementation starts by ingesting three structured clinical datasets, PEP-T1, PEP-T2, and PEP-T3, corresponding to first-, second-, and third-trimester measurements, respectively. Each dataset contains 1000 patient records, each uniquely identified by a PublicID, and annotated with a binary outcome variable (result): OUTCOME=0 indicates the absence of preeclampsia, and OUTCOME=1 indicates the presence of preeclampsia. PEP-T1 includes 56 first-trimester (11-14 weeks) features and focuses on biochemical markers such as ADAM12, Endoglin, sFlt-1, VEGF, AFP, fbHCG, Inhibin-A, and PAPP-A. PEP-T2 includes 99 features measured in the second trimester (18-22 weeks), whereas PEP-T3 includes 144



International Journal of
Medical Toxicology & Forensic Medicine

Figure 3. Structural design of the proposed PEP-Ensem-2N framework.

features, representing clinical indicators and biochemical markers, in the third trimester (28-32 weeks).

$$X^* = \text{Unique}(X) \quad (1)$$

After data acquisition, we preprocess the dataset. Partitioning the data into trimesters enables longitudinal characterization of the disease trajectory, allowing the model to capture both early biochemical perturbations and late-stage clinical manifestations. We remove duplicate entries by keeping only unique patient instances, as in Eq. (1). This precaution prevents artificial inflation of performance metrics and ensures each patient contributes individually to learning.

$$x_{i,j}^{imp} = \frac{1}{|\Omega_j|}\sum_{k\in\Omega_j} x_{k,j}^{obs} \quad (2)$$

Clinical datasets often have missing data due to lab measurement errors or because certain examinations are unavailable. Instead of discarding these samples, which might reduce statistical power, we use column-wise mean imputation. Let Ω_j be the set of samples in which feature j is observed. Mean imputation preserves the statistical distribution of each feature and dataset completeness, which is particularly important when the number of positive cases is small, as shown in Eq. (2).

$$y_i \in \{0,1\} \quad (3)$$

The categorical outcome variable is represented as a binary classification, $y_i = 1$ for preeclampsia and $y_i = 0$ for non-preeclampsia, according to Eq. (3). This representation makes direct optimization with binary classification loss functions easy.

$$X \in R^{N\times m}, \quad y \in \{0,1\}^N \quad (4)$$

After encoding, the dataset takes the form of a predictor matrix X∈R^ {N×m}, and a target vector y∈ {0,1} ^N, where N represents the number of samples, and \(m\) represents the number of input features. This numerical encoding enables direct optimization with binary cross-entropy loss and easy interfacing with deep neural networks, as described in Eq. (4).

$$x'_{i,j} = \frac{x_{i,j}-\mu_j}{\sigma_j} \quad (5)$$

Clinical variables show heterogeneous ranges (e.g., hormone concentrations, blood pressure, biochemical levels). To reduce scale effects, we apply z-score normalization according to Eq. (5), where μj and σj are the mean and standard deviation of feature j, respectively. This transformation ensures each feature has a zero mean and unit variance, stabilizing gradient descent and accelerating convergence. Through systematic cleaning, imputation, normalization, stratified partitioning, multi-factor oversampling, class

re-weighting, and tensor reshaping, the pre-processing stage builds a noise-reduced, balanced, and numerically stable dataset. Standardization equalizes feature magnitudes, prevents large-scale variables from dominating, and conditions the optimization landscape to support faster convergence through more stable gradient descent.

$$\left(X_{train}^{aug}, y_{train}^{aug}\right) = \text{SMOTE}(X_{train}, y_{train}) \quad (6)$$

Unlike conventional single-pass oversampling, SMOTE is applied only to the training set to generate augmented samples, which helps avoid leaking information into the test data. This multi-factor augmentation causes the model to see different variations of the minority class, and the class separability of the training data is greatly improved, and sampling bias is also reduced as required in Eq. (6). Thereafter, the standardized data set is divided into a training and a testing set with an 80:20 ratio.

$$I_j = \frac{1}{T}\sum_{t=1}^{T} \Delta i_j^{(t)} \quad (7)$$

A Random Forest classifier consists of 200 trees that estimate feature importance using the mean decrease in impurity criterion (Eq. (7)). The Random Forest-based importance ranking helps identify the most discriminative predictors and suppress noisy and redundant attributes, leading to a compact but informative feature representation.

$$F_{(top)} = argmax_{(|F|=15)} I_j \quad (8)$$

Features are ranked by their importance scores Ij, and the top fifteen features are kept, as described in Eq. (8). To model complex temporal patterns in longitudinal maternal clinical data, we use various deep learning architectures. Recurrent models such as Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) networks capture sequential dependencies. In contrast, Temporal Convolutional Networks (TCNs) are added to model long-range temporal relationships efficiently. Moreover, we add a Deep Convolutional Network (DCN) to extract hierarchical feature representations. We then integrate these models' complementary strengths using an ensemble strategy to improve predictive robustness. The LSTM unit's internal gating operations, such as the forget gate, input gate, and candidate cell state, are defined in Eqs. (9-11).

$$f_t = \sigma\left(W_f[h_{t-1}, x_t] + b_f\right) \quad (9)$$

$$i_t = \sigma(W_i[h_{t-1}, x_t] + b_i) \quad (10)$$

$$\widetilde{C_t} = \tanh(W_c[h_{t-1}, x_t] + b_c) \quad (11)$$

$$C_t = f_t \odot C_{t-1} + i_t \odot \widetilde{C_t} \quad (12)$$

$$o_t = \sigma(W_o[h_{t-1}, x_t] + b_o) \quad (13)$$

$$h_t = o_t \odot \tanh(C_t) \quad (14)$$

The Long Short-Term Memory (LSTM) module uses gating mechanisms to control the flow of information from one-time step to the next. Sigmoid activation functions are used in the forget, input, and output gates, while the hyperbolic tangent function is used in the candidate cell state to bound the memory values and maintain stable learning data. (Eq.12) Overall, these are the operations that update the current cell state, and it involves a balancing of memory retention and incorporation of new data. Thereafter, the output gate (Eq. 13) selects the contribution of the cell state to the updated hidden representation h in Eq. (14), thereby facilitating the modeling of long-term temporal patterns. The gating mechanisms of the Gated Recurrent Unit (GRU) architecture, i.e., the update gate, reset gate, candidate activation, and hidden state update, are defined as follows in Eqs. (15)-(18).

$$z_t = \sigma(W_z[h_{t-1}, x_t] + b_z) \quad (15)$$

$$r_t = \sigma(W_r[h_{t-1}, x_t] + b_r) \quad (16)$$

$$\widetilde{h_t} = \tanh(W_h[r_t \odot h_{t-1}, x_t] + b_h) \quad (17)$$

$$h_t = (1 - z_t) \odot h_{t-1} + z_t \odot \widetilde{h_t} \quad (18)$$

The GRU module simplifies recurrent computation by combining memory and hidden states and maintaining effective temporal modeling through update and reset gates implemented with sigmoid activations. The update gate in Eq. (15) mediates the contributions made by prior and current states, while the reset gate in Eq. (16) ascertains how far historical information is to be discarded. The candidate activation in Eq. (17) combines the reset gate, the previous hidden state, and the current input, and the hidden state update in Eq. (18) selectively learns relevant dependencies and filters out redundant patterns.

$$h_t = \tanh(W_h h_{t-1} + W_x x_t + b) \quad (19)$$

Complementing these, there is a light-weight mechanism for temporal representation - where the hidden state h_t in Eq. (19) changes according to preceding states, current inputs and associated weights, is provided by the PEP - Simple RNN learner. In this learner, the Simple RNN uses a hyperbolic tangent activation function to restrict the hidden state to the range [-1,1], making temporal learning more stable and easier to learn nonlinear temporal relationships.

Collectively, this multi-stream ensemble offers an efficient strategy for integrating deep recurrent and convolutional architectures to create accurate, robust models of multi-temporal clinical data for early preeclampsia risk prediction. Gated architectures selectively retain important clinical temporal dependencies and suppress noise, enabling modeling of changing biomarker trajectories across trimesters. Simultaneously, the DCN branch uses stacked one-dimensional convolutional layers with an increasing number of filters (50 and 100) and a kernel size of three in order to extract fine-grained sequential dependencies. These layers use ReLU, L2 Normalization, and then global max pooling, dropout, and further normalization to retain dominant temporal signals and reduce noise. The temporal convolution operations used in the TCN branch, causal and dilated convolutions that model long-range dependencies, are defined in Eqs. (20) and (21).

$$h_t = \sum_{k=0}^{K-1} w_k x_{t-k} \quad (20)$$

Meanwhile, the TCN stream includes hierarchical dilated causal convolutions and dense units to effectively model multi-scale temporal dependencies. It starts with a hidden layer of 100 ReLU Neurons with L2 Regularization, Dropout, and dual normalization, and ends with a dense layer of 50 ReLU Neurons that creates high-level temporal abstractions that are complementary to the recurrent and convolutional learners. The PEP architecture incorporates causal convolution to preserve temporal order, so the output at time step t, denoted h_t, depends only on the current and past inputs, as depicted in the PEP-TCN model. Mathematically, this causal relationship is expressed in Eq. 20, where w_k represents the convolutional filter weights, K is the kernel size, and x_{t-k} denotes the input at previous time steps ensuring the causal dependence.

$$h_t = \sum_{k=0}^{K-1} w_k x_{t-dk} \quad (21)$$

To efficiently expand the temporal receptive field, we use dilated convolution with dilation factor d, which skips input samples at predefined intervals. As defined in Eq. (21), dilated convolution adds a dilation factor d, which distances the convolutional kernel from previous inputs, which therefore allows to effectively model long-range temporal dependencies.

$$Z_{l+1} = \mathrm{ReLU}(W_l Z_l + b_l) \quad (22)$$

By increasing the dilation rates, the Temporal Convolutional Network (TCN) can capture patterns at different time scales, making it powerful for modeling complex sequential dynamics. Concomitantly, the PEP-DCN branch builds a hierarchical feature representation using stacked convolutional layers, each applying a weighted transformation followed by a nonlinear activation. The transformation at layer i+1 is given by Eq. (22), where W_l and b_l are the weight matrix and bias vector of the l-th convolutional layer, respectively. In this setup, the rectified linear unit (ReLU), denoted by f, serves as the nonlinear activation function, enabling deeper layers to drive higher-level learning while lower layers focus on lower-level pattern extraction.

$$P_{ensemble} = \sum_{i=1}^{N} w_i P_i \quad (23)$$

Finally, we concatenate the output sequences (LSTM, GRU, Simple RNN, TCN, and DCN) using a weighted sum layer, multiplying each sequence by its corresponding ensemble coefficient. Next, after a dense layer with sigmoid activation, we obtain a concatenated representation that is transformed into a binary probability for "Preeclampsia yes" or "Preeclampsia no." The final ensemble probability is obtained by weighted soft voting, where P_i is the predicted probability of the i-th base learner, and w_i is the corresponding weight, given by Eq. (23). With N representing the total number of models, w_i is the weight of the i-th model, and P_i is the predicted probability for the i-th model. In the current implementation, the weights are defined as: SimpleRNN = 0.2w, GRU = 0.2w, LSTM = 0.2w, TCN = 0.2w, and DCN = 0.1w, so that their sum contributes in a balanced manner based on the positive proficiency in predictions. To maintain architectural consistency, the inputs to the three-dimensional models (Simple RNN, GRU, LSTM and TCN) are reformatted as three-dimensional tensors, while the fully connected deep neural network (DNN) takes the input in its original two-dimensional form.

$$\hat{y} = \begin{cases} 1, & if\ p \\ 0, & \end{cases} \text{ensemble} > 0.5 \quad (24)$$

The final class prediction is obtained by applying a decision threshold, as shown in Eq. 24. Optimization uses the Adam optimizer with a learning rate of 0.001 and binary cross-entropy loss to ensure stable convergence. To overcome overfitting, early stopping with patience = 10 epochs is used, coupled with a ReduceLROnPlateau scheduler (factor = 0.5, patience = 5, min-delta = 10-6) to achieve adaptive learning rate adjustments.

$$\mathrm{ReLU}(x) = \max(0, x) \quad (25)$$

In the PEP implementation, the ReLU activation function given in Eq. (25) is used in the hidden layers to introduce nonlinearity.

$$\sigma(z) = \frac{1}{1+e^{-z}} \quad (26)$$

The sigmoid activation function outlined in Eq. (26) is used in the output layer to project network outputs into probabilities in the interval [0,1]. Collectively, the PEP-Ensem-2N implementation unifies rigorous preprocessing, multi-factor augmentation, Random Forest-driven feature refinement, multi-architecture deep learning and probabilistic ensemble fusion under a coherent framework. This design provides a scalable, generalizable, and clinically meaningful solution for early and accurate preeclampsia prediction.

Results

The Random Forest-based feature-importance approach provides an interpretable framework for identifying clinically salient predictors, helping clinicians understand which biomarkers and clinical variables most influence preeclampsia risk estimation in each trimester. Drawing on empirical findings from the implementation phase, this paper reports a detailed analysis of the experimental results and compares the predictive performance of the proposed PEP-Ensem-2N model with conventional baselines, highlighting its robustness and clinical relevance. Figure 4 shows the Random Forest-derived feature-importance rankings for the top 15 features in the PEP-T1, PEP-T2, and PEP-T3 datasets (Figure 3).

Each histogram shows how feature contributions differ across the three trimesters, with the most influential variables on the left. These visualizations support a nuanced, relative interpretation of each feature's impact on model performance. Across the three datasets, several features, such as V1BA0631, the MAP-derived metrics, and several V2BA and UxC coded clinical indicators, consistently appear as the leading predictors, demonstrating their stability and robustness in predicting preeclampsia.

$$L_{BCE} = -[y_i \log(p_i) + (1 - y_i) \log(1 - p_i)] \quad (27)$$

The binary cross-entropy loss, as defined in Eq. (27), penalizes incorrect preeclampsia predictions and guides model optimization.

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} \quad (28)$$

$$\text{Precision} = \frac{TP}{TP+FP} \quad (29)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (30)$$

$$\text{F1} = \frac{2\,\text{Precision}\,\text{Recall}}{\text{Precision}+\text{Recall}} \quad (31)$$

Equation (28) defines accuracy as the ratio of correctly classified data instances to the total number of data instances, reflecting the model's overall predictive accuracy. Equation (29) defines precision, the reliability of positive predictions, as the percentage of correctly identified positive preeclampsia cases among all predicted positives. Equation (30) defines recall (or sensitivity), which measures how well the model identifies true preeclampsia cases and avoids missing any. Equation (31) gives the F1-score, the harmonic mean of precision and recall, and thus a balanced measure of classification performance, especially in imbalanced classification tasks.

The epoch-wise training and validation statistics for the first ten epochs of the PEP-Ensem-2N model are given in Tables 1-3; these tables describe the learning dynamics, convergence characteristics, and stability of the model on the PEP-T1, PEP-T2, and PEP-T3 datasets.

For the PEP-Ensem-2N setup with the ReLU activation function, the tables show metrics such as accuracy, step loss, validation accuracy, and validation loss. The reported metrics show early-stage learning trends and indicate that training converges over the full

Table 1. Epoch history of the PEP-T1-PEP-Ensem-2N model.

Epoch	Step Loss	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
1/55	0.0928	0.0327	0.0781	0.2453	0.5380
2/55	0.4071	0.8214	0.0348	0.5928	0.7466
3/55	0.3709	0.8346	0.3590	0.5210	0.0485
4/55	0.7138	0.8923	0.3634	0.9679	0.2508
5/55	0.5706	0.9517	0.7565	0.1581	0.1055
6/55	0.9643	0.8307	0.2323	0.9555	0.7207
7/55	0.7123	0.2194	0.4868	0.9888	0.6230
8/55	0.5021	0.6346	0.6412	0.7117	0.0909
9/55	0.3913	0.9083	0.5084	0.3186	0.2892
10/55	0.1336	0.9702	0.0066	0.2927	0.9681

Table 2. Epoch history of PEP-T2-PEP-Ensem-2N model.

Epoch	Step Loss	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
1/55	0.4982	0.7530	0.3558	0.2755	0.9635
2/55	0.0724	0.5495	0.1678	0.0197	0.2066
3/55	0.7481	0.8180	0.7406	0.5997	0.9046
4/55	0.8884	0.8089	0.2553	0.8565	0.1877
5/55	0.6361	0.6481	0.7165	0.5794	0.2665
6/55	0.4827	0.3734	0.3148	0.9714	0.0021
7/55	0.7823	0.9941	0.5162	0.1970	0.0594
8/55	0.4066	0.6111	0.1841	0.1254	0.1383
9/55	0.1915	0.5802	0.9851	0.8883	0.3285
10/55	0.1107	0.0692	0.3719	0.1984	0.2832

International Journal of
Medical Toxicology & Forensic Medicine

Table 3. Epoch history of PEP-T3-PEP-Ensem-2N model.

Epoch	Step Loss	Training Accuracy	Training Loss	Validation Accuracy	Validation Loss
1/55	0.5214	0.8741	0.4127	0.8126	0.4389
2/55	0.3876	0.9128	0.3254	0.8653	0.3015
3/55	0.2989	0.9436	0.2611	0.9034	0.2248
4/55	0.2145	0.9659	0.1983	0.9317	0.1769
5/55	0.1623	0.9784	0.1472	0.9542	0.1247
6/55	0.1184	0.9861	0.1095	0.9728	0.0912
7/55	0.0836	0.9917	0.0784	0.9849	0.0648
8/55	0.0591	0.9952	0.0569	0.9913	0.0427
9/55	0.0418	0.9979	0.0397	0.9961	0.0294
10/55	0.0286	0.9990	0.0261	0.9990	0.0189

International Journal of
Medical Toxicology & Forensic Medicine

55 epochs. We then extend the comparison with a detailed analysis of epoch-wise convergence, inter-model variability, and performance differences attributable to the activation function, providing a strong foundation for claiming the superiority of the PEP-Ensem-2N architecture over the alternative models.

Table 4 lists the performance of several models, namely PEP-Ensem-2N, PEP-LSTM, PEP-GRU, PEP-DCN, PEP-Simple-RNN, and PEP-TCN, across the three trimesters (T1, T2, T3) and with various activation functions (ReLU, Sigmoid, Tanh, and Softmax). The reported accuracy values include all combinations of model, activation function, and trimester-specific dataset, revealing performance variations due to both activation functions and dataset partitions. The ReLU-based PEP-Ensem-2N achieves 0.9990 accuracy on PEP-T3 and near-perfect accuracy on PEP-T1 (0.9946) and PEP-T2 (0.9992), making it the best-performing model among all model-activation configurations on the evaluated datasets.

Building on these results, the following analysis examines the proposed model's predictive stability and relative merits. Figures 5-8 show the training and validation learning curves of six models, PEP-Simple-RNN, PEP-GRU, PEP-LSTM, PEP-TCN, PEP-DCN, and PEP-Ensem-2N, on the PEP-T1, PEP-T2, and PEP-T3 datasets. The first two figures show the training accuracy and training loss curves, respectively, while the last two figures show the validation accuracy and validation loss curves, respectively, which provide insight into each model's optimization process and generalization ability. This study did not include an independent external validation cohort. Therefore, the reported results should not be interpreted as transportable to different institutions, populations, or clinical settings. Interpret the evaluation during pregnancy visits in the context of the study's data source and sampling design. The framework needs external validation using independently collected clinical observations and assessment of calibration and clinical utility (Figure 9).

Table 4. Accuracy for all models.

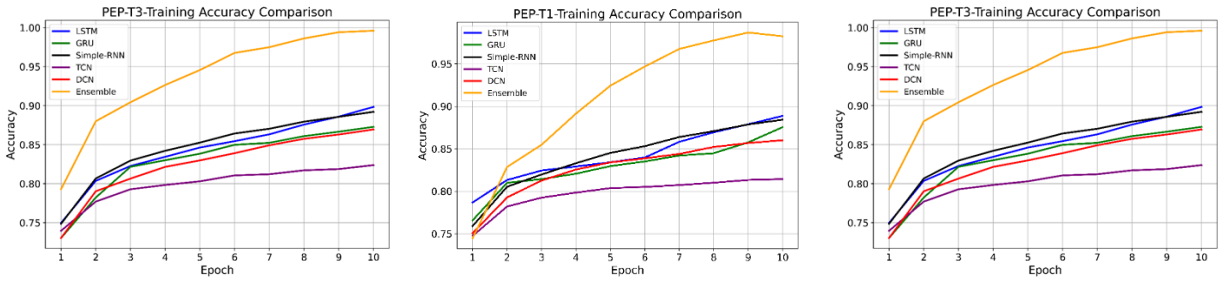
Model Name	Activation Function	Accuracy		
		PEP-T1	PEP-T2	PEP-T3
PEP-Ensem-2N	Relu	0.9946	0.9992	0.9990
	Sigmoid	0.8390	0.1740	0.8263
	Tanh	0.9923	0.8260	0.9990
	Softmax	0.8350	0.8263	0.1756
PEP-LSTM	Relu	0.3470	0.4300	0.7296
	Sigmoid	0.8356	0.1740	0.1806
	Tanh	0.7340	0.8260	0.1740
	Softmax	0.1650	0.8260	0.3236
PEP-GRU	Relu	0.8336	0.2490	0.1750
	Sigmoid	0.8350	0.8260	0.8256
	Tanh	0.7387	0.8260	0.7490
	Softmax	0.8350	0.8260	0.1740
PEP-Simple-Rnn	Relu	0.9783	0.9870	0.9890
	Sigmoid	0.8360	0.1740	0.8263
	Tanh	0.9730	0.7300	0.9990
	Softmax	0.8350	0.7310	0.8260
PEP-DCN	Relu	0.4933	0.6696	0.5173
	Sigmoid	0.1650	0.8260	0.8260
	Tanh	0.6020	0.7390	0.6956
	Softmax	0.8350	0.7700	0.1740
PEP-TCN	Relu	0.5896	0.7160	0.4670
	Sigmoid	0.8356	0.1740	0.8030
	Tanh	0.3923	0.7600	0.7730
	Softmax	0.1650	0.1750	0.1740

International Journal of Medical Toxicology & Forensic Medicine

Discussion

This study examined an ensemble model (PEP-Ensem-2N) that first selected features and then modeled PEP risk at pregnancy visits, using a class-imbalance ratio. The authors considered an ensemble model (PEP-Ensem-2N) that first selected features and then modeled PEP risk at pregnancy visits, using a class-imbalance ratio. The reported accuracies of the rectified linear unit-based ensemble were 0.9946, 0.9992, and 0.9990 for PEP-T1, PEP-T2, and PEP-T3, respectively. The values show good classification performance in the reported experiments, but are valid only if the data partitions, preprocessing methods, and evaluation denominators are consistent. These should therefore be considered preliminary until the inconsistencies in the current analysis have been clarified. We combine recurrent and convolutional models because they may capture complementary relationships between input variables that different architectures can learn. Weighted probability fusion can reduce the risk of errors in a single model. However, it does not benefit individual learners in an ensemble with a variety of architectures. Compare the ensemble's performance with its constituent models using the same voting method as equal-weight voting, as well as with appropriate conventional baselines under the same evaluation conditions. Also note that the

weighting should be explained: possible limitations to weights, normalization, etc. of weights. In addition, three-dimensional tensorization of the tabular predictors does not capture changes during pregnancy along the longitudinal dimension. Major impediments of the present evaluation are data provenance and sample accounting. This is because the three confusion matrices report different numbers of observations (1000) and participants (5010, 4956, and 4956). Also, some counts differ from the description. These mismatches compromise the ability to correctly interpret the evaluation sample, and they impact the understanding of unique participants, visit records, augmented training observations and repeated prediction events. All of the referenced PEPrML's sample files in the public repository are synthetic demonstration data. Using these files, performance would be calculated from observations in these files, not predictive performance within the observed pregnancy cohort.



International Journal of Medical Toxicology & Forensic Medicine

Figure 5. Training accuracy for all datasets and all PEP-T models.

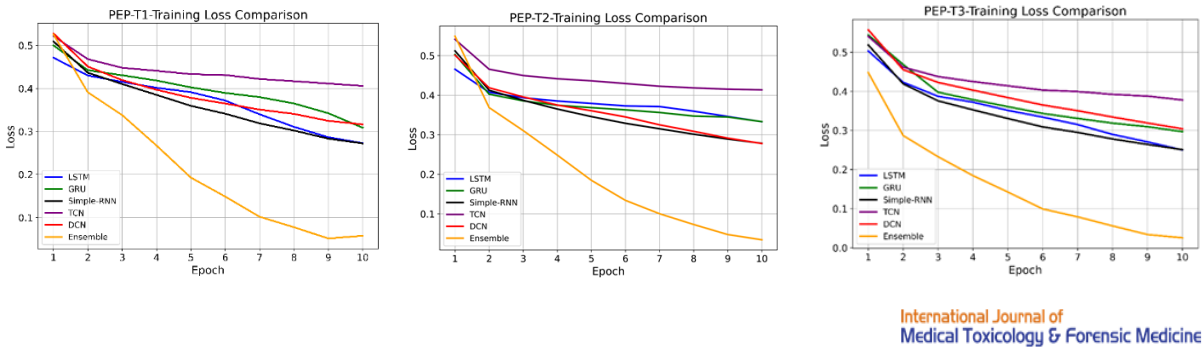


Figure 6. Training loss for all dataset and all PEP-T models.

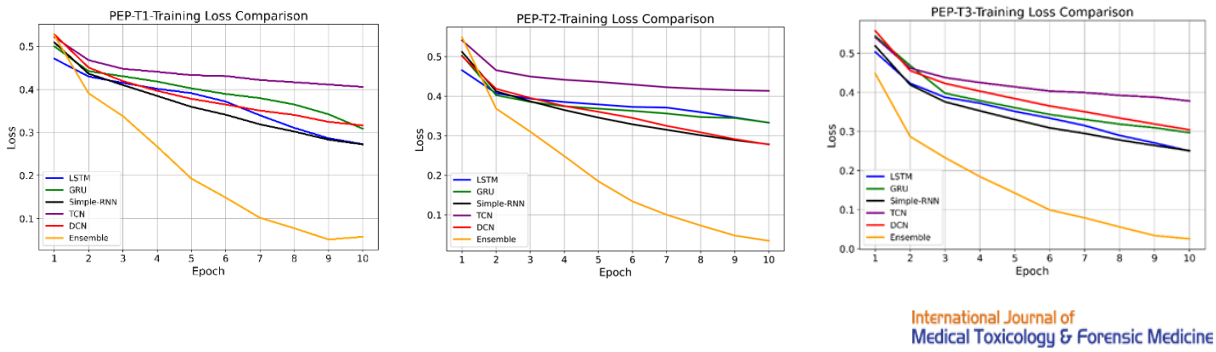


Figure 7. Validation accuracy for all dataset and all PEP-T models.

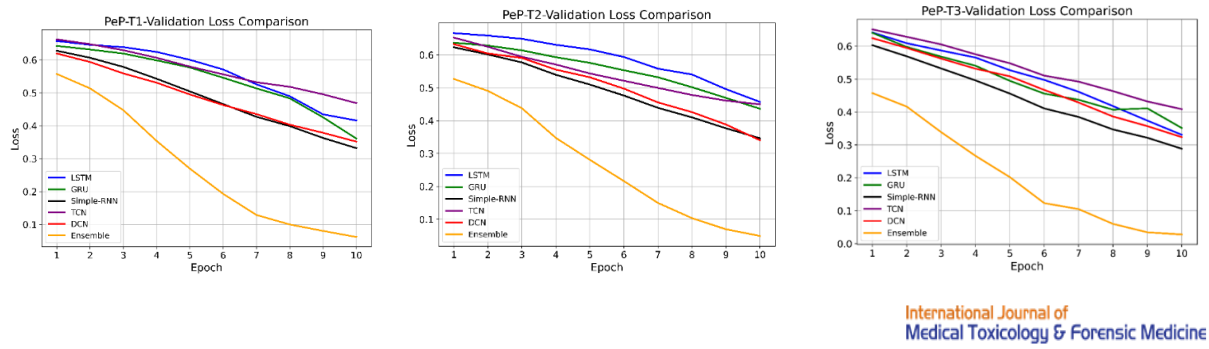


Figure 8. Validation loss for all datasets and all PEP-T models.

This close match in accuracy further highlights the need to eliminate information leakage and implementation errors. Imputation, standardization, feature selection, and resampling procedures should aim to be done within the training partitions (even during CV). Store each record in the same partition, and ensure predictors are available at prediction time. Care is also needed when interpreting synthetic minority oversampling. It can tolerate unbalanced class frequencies but may violate relationships among clinical measurements, categorical variables, and

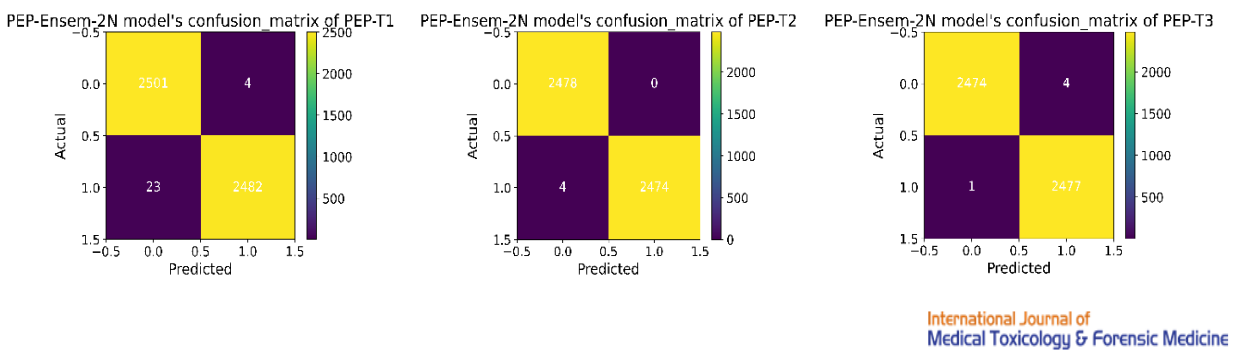


Figure 9. Confusion matrix of PEP-Ensem-2N with all PEP-T datasets.

derived biomarker ratios in the generated data. The contribution should then be compared with approaches that do not resample and use class weighting. Existing work reports imbalanced outcome frequencies when predicting preeclampsia using imbalance-aware neural networks [11]. The efficiency differences among activation functions need to be explored. Some of the activation functions may be used improperly, and the values of accuracy close to 0.17-0.18 should not be considered representative of this. Possible issues include predictions in a single class, label-encoding differences, mismatches between the output layer and loss, or mislabeled probabilities. The conclusions therefore depend on checking prediction distributions, matching labels, and calculating output dimensions and measures. Feature importance describes the contribution of selected variables in a model to prediction. It is not used to imply a causal interpretation in terms of biological mechanisms or the validity of the biomarkers or clinical relevance. Variables with the same name in the computational dataset should not be replaced with the biomarker ratios and angiogenic factors studied in the mother [16, 17]. Likewise, differences in populations studied, outcome definitions, prediction timing, and validation procedures limit direct comparison with other studies [2-4, 19-20]. Last but not least, a clinical prediction model is not just about accuracy. It needs to be sensitive and specific, with positive and negative predictive values to describe its strengths and limitations, and it should include discrimination, calibration, and confidence intervals. If clinically relevant information and clinically meaningful decision thresholds are available, decision-curve analysis can provide additional information. This study had no external independent validation, so generalisability to other institutions or patient groups cannot be guaranteed. Future research should further evaluate PEP-Ensem-2N for clinical decision support by improving transparent data provenance, ensuring reproducible evaluation, and conducting independent clinical validation.

Conclusion

This work presents PEP-Ensem-2N, a unified deep ensemble framework for accurate, reliable preeclampsia prediction in trimester-specific clinical datasets. By combining multi-factor SMOTE-based augmentation, Random Forest-based feature optimization, and heterogeneous-driven deep learning architectures in a weighted soft-voting ensemble, the proposed system addresses major issues such as class imbalance, high dimensionality, and non-linear trends in disease patterns. Comprehensive evaluation shows that the ReLU-based PEP-Ensem-2N consistently outperforms the individual deep learning models and other activation configurations, achieving near-perfect predictive accuracy on the PEP-T1, PEP-T2, and PEP-T3 datasets. Stable training and validation curves and highly favorable confusion matrix results confirm the framework's potential for robust generalization. Moreover, identifying common influential features across trimesters supports the model's clinical interpretability and its potential to assist biomarker-driven decision-making. Overall, PEP-Ensem-2N is a robust, scalable, and clinically meaningful strategy for early preeclampsia risk assessment, with the potential to facilitate timely risk assessment and early intervention, as well as personalized monitoring and improved maternal-fetal outcomes.

Ethical Approval

Not applicable.

Acknowledgment

None.

Funding

No external funding supported this research.

Conflicts of Interest

The authors report no competing interests to declare.

References

[1] Edvinsson C, Björnsson O, Erlandsson L, Hansson SR. Predicting intensive care need in women with preeclampsia using machine learning – a pilot study. Hypertens Pregnancy. 2024;43(1):2312165. [DOI: 10.1080/10641955.2024.2312165]

[2] Torres-Torres J, Villafan-Bernal JR, Martinez-Portilla RJ, Hidalgo-Carrera JA, Estrada-Gutierrez G, Adalid-Martinez-Cisneros R, et al. Performance of machine-learning approach for prediction of pre-eclampsia in a middle-income country. Ultrasound Obstet Gynecol. 2024;63(4):506-13. [DOI: 10.1002/uog.27510]

[3] Li T, Xu M, Wang Y, Wang Y, Tang H, Duan H, et al. Prediction model of preeclampsia using machine learning based methods: a population-based cohort study in China. Front Endocrinol (Lausanne). 2024;15:1345573. [DOI: 10.3389/fendo.2024.1345573]

[4] Butler L, Gunturkun F, Chinthala L, Karabayir I, Tootooni MS, Bakir-Batu B, et al. AI-based preeclampsia detection and prediction with electrocardiogram data. Front Cardiovasc Med. 2024;11:1360238. [DOI: 10.3389/fcvm.2024.1360238]

[5] Scripcariu IS, Vasilache IA, Pavaleanu I, Doroftei B, Cărăuleanu A, Socolov D, et al. Machine Learning-Based Prediction of Intrauterine Growth Restriction and Preeclampsia: A Prospective Study. Med Sci (Basel). 2024;12(1):8. [DOI: 10.3390/medsci12010008]

[6] Villalaín C, Herraiz I, Domínguez-Del Olmo P, Angulo P, Ayala JL, Galindo A. Prediction of Delivery Within 7 Days After Diagnosis of Early Onset Preeclampsia Using Machine-Learning Models. Front Cardiovasc Med. 2022;9:910701. [DOI: 10.3389/fcvm.2022.910701]

[7] Gomez-Jerez L, Oprescu AM, Chimenea-Toscano Á, Garcia-Diaz L, Romero-Ternero MDC. Machine Learning to Predict Pre-Eclampsia and Intrauterine Growth Restriction in Pregnant Women. Electronics. 2022;11(19):3240. [DOI: 10.3390/electronics11193240]

[8] Chen X, Lan Y, Du J, Lin W, Ke J, Lin F, et al. Development and validation of the prediction models for preeclampsia: a retrospective, single-centre, case-control study. Ann Transl Med. 2022;10(18):987. [DOI: 10.21037/atm-22-4192]

[9] Zhang X, Chen Y, Salerno S, Li Y, Zhou L, Zeng X, et al. Prediction of severe preeclampsia in machine learning. Med Novel Technol Devices. 2022;14:100158. [DOI: 10.1016/j.medntd.2022.100158]

[10] McLaughlin K, Hobson SR, Ravi Chandran A, Agrawal BS, Windrim RC, Parks WT, et al. Circulating maternal placental growth factor responses to low-molecular-weight heparin in pregnant patients at risk of placental dysfunction. Am J Obstet Gynecol. 2022;226(2):248.e1-248.e12. [DOI: 10.1016/j.ajog.2021.08.027]

[11] Bennett R, Mulla ZD, Parikh P, Hauspurg A, Razzaghi T. An imbalance-aware deep neural network for early prediction of preeclampsia. PLoS One. 2022;17(3):e0266042. [DOI: 10.1371/journal.pone.0266042]

[12] Gupta K, Balyan K, Lamba B, Puri M, Sengupta D, Kumar M. Ultrasound placental image texture analysis using artificial intelligence to predict hypertension in pregnancy. J Matern Fetal Neonatal Med. 2022;35(25):5281-7. [DOI: 10.1080/14767058.2021.1887847]

[13] Wang G, Zhang Y, Li S, Zhang J, Jiang D, Li X, et al. A Machine Learning-Based Prediction Model for Cardiovascular Risk in Women with Preeclampsia. Front Cardiovasc Med. 2021;8:736491. [DOI: 10.3389/fcvm.2021.736491]

[14] Zhang Y, Wang X, Han N, Zhao R. Ensemble Learning Based Postpartum Hemorrhage Diagnosis for 5G Remote Healthcare. IEEE Access. 2021;9:20546-20556. [DOI: 10.1109/ACCESS.2021.3051215]

[15] Li YX, Shen XP, Yang C, Cao ZZ, Du R, Yu MD, et al. Novel electronic health records applied for prediction of pre-eclampsia: Machine-learning algorithms. Pregnancy Hypertens. 2021;26:104-110. [DOI: 10.1016/j.preghy.2021.10.006]

[16] Huang Q, Hao S, You J, Yao X, Li Z, Schilling J, et al. Early-pregnancy prediction of risk for pre-eclampsia using maternal blood leptin/ceramide ratio: discovery and confirmation. BMJ Open. 2021;11(9):e050963. [DOI: 10.1136/bmjopen-2021-050963]

[17] Dathan-Stumpf A, Czarnowsky V, Hein V, Andraczek T, Stepan H. Real-world data on the clinical use of angiogenic factors in pregnancies with placental dysfunction. Am J Obstet Gynecol. 2021;224(4):393.e1-393.e11. [DOI: 10.1016/j.ajog.2020.10.028]

[18] Sufriyana H, Wu YW, Su EC. Prediction of Preeclampsia and Intrauterine Growth Restriction: Development of Machine Learning Models on a Prospective Cohort. JMIR Med Inform. 2020;8(5):e15411. [DOI: 10.2196/15411]

[19] Marić I, Tsur A, Aghaeepour N, Montanari A, Stevenson DK, Shaw GM, et al. Early prediction of preeclampsia via machine learning. Am J Obstet Gynecol MFM. 2020;2(2):100100. [DOI: 10.1016/j.ajogmf.2020.100100]

[20] Jhee JH, Lee S, Park Y, Lee SE, Kim YA, Kang SW, et al. Prediction model development of late-onset preeclampsia using machine learning-based methods. PLoS One.

2019;14(8):e0221202. [DOI: 10.1371/journal.pone.0221202]

[21] PRAISE Lab. PEPrML: Machine Learning for Preeclampsia Prediction [Software/Repository]. GitHub; 2023 [cited 2024]. [Link]