Ovarian Cancer Prediction: A Stacking Ensemble Approach

Marilena Marmora, Michele Roccotelli, *Senior Member, IEEE*, Wasim Ali, Maria Pia Fanti, *Fellow, IEEE*, Luis Carlos Félix-Herrán, *Member, IEEE*

***Abstract*—Ovarian cancer is one of the most aggressive neoplasms worldwide, representing the eighth leading cause of cancer death in women, having a survival rate ranging from 90% for early diagnosis to 13.4% for advanced stages. This study aims to propose a reliable and accurate diagnostic support system via the development of Machine Learning models — Random Forest, Support Vector Machine, Decision Tree, Simple Artificial Neural Network and XGBoost — based exclusively on routine clinical and hematological parameters, characterised by low cost and easy availability. The study used a unified dataset of 933 samples and 48 common features, previously preprocessed using a hybrid imputation technique to handle missing values, improving performance compared to the reference study. The core of this work lies in the implementation of the Stacking Ensemble model with a Logistic Regression-based Meta Learner, which enhances the performance achieved by the most promising Base Learner. Experimental results demonstrate that the Stacking Ensemble outperforms the best individual models, achieving an accuracy of 0.9840 and a specificity of 1.0000. This evidence confirms the potential of Artificial Intelligence as a concrete support tool to improve the reliability of clinical decisions and early diagnosis.**

***Index Terms*—ovarian cancer, hematological markers, machine learning, stacking ensemble, hybrid imputation**

I. Introduction

Among the malignancies affecting the female reproductive system, ovarian cancer — defined in the literature as the "silent killer" — has played a critical role in medical research in recent years due to the severe diagnostic challenges arising from the clinical heterogeneity with which the disease manifests and the absence of specific symptoms that severely hinders early diagnosis. Currently, it represents the eighth leading cause of cancer death in women globally: data report a 5-year survival rate exceeding 90% in women with Stage I disease; however, since approximately 70% of cases are diagnosed when the disease has reached Stage III or IV, the survival rate drops significantly, reaching 13.4% at Stage IV [1].

The disease manifests itself in various histotypes with different levels of aggressiveness and is influenced by a complex interaction among genetic factors (such as BRCA1 and BRCA2 germline mutations), hormonal factors related to the use of contraceptives or hormone replacement therapy and lifestyle factors such as obesity and physical activity [2]. This clinical heterogeneity, coupled with the lack of specific early symptoms, hampers early diagnostic options and the development of focused prevention strategies.

The current diagnostic protocols integrate imaging techniques and biomarker monitoring. Transvaginal ultrasound, despite being the primary imaging tool, is often insufficient for accurately distinguishing between benign and malignant lesions in early stages [3]. Advanced techniques such as Magnetic Resonance Imaging and Computed Tomography ensure higher accuracy [4]; nevertheless, they entail high costs, require highly qualified specialists and are unable to autonomously identify the stage or histological subtype. Concurrently, routine serum markers — primarily Carbohydrate Antigen 125 (CA-125) — are widely adopted due to its easy detection. However, elevated CA-125 levels do not always indicate malignancy, since a physiological increase in CA-125 can also be associated with noncancerous conditions such as cysts or inflammation. As a consequence, definitive diagnosis still requires the use of invasive procedures (e.g. histological biopsy).

In this context, Artificial Intelligence (AI), with Machine Learning (ML) and Deep Learning (DL) algorithms [5], [6], has opened new perspectives in medicine and in disease diagnosis via the integrated analysis of radiological images and histopathology and genomic data. These technologies facilitate the recognition of complex patterns that are not easy to identify with traditional clinical methods. In recent years, the application of AI in ovarian cancer diagnosis has been at the center of numerous studies, which have primarily focused on the use of preoperative data and biomarkers. Akazawa and Hashimoto demonstrated the effectiveness of the XGBoost algorithm, achieving 80% accuracy [7]. A significant contribution was recently provided by Ayyoubzadeh et al. [8], who analysed a cohort of 349 records. Using single classifiers such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM) and Artificial Neural Networks (ANN), the study identified Human Epididymis Protein 4 (HE4), CA-125 and Neutrophil Ratio (NEU) as the most relevant predictive factors, achieving an accuracy of 86.75% with the RF model. However, the authors emphasized that the generalizability of the results is limited by the small size and the data's origin from a single center. In the context of image analysis, Paayas and Annamalai implemented the DenseNet-121 architecture on histopathological images, achieving excellent discriminatory results, with an accuracy of 99.6% and an F1-score of 1.00 [9]. However, the application of data augmentation techniques

Marilena Marmora (m.marmora1@studenti.poliba.it), Michele Roccotelli (michele.roccotelli@poliba.it), Wasim Ali (wasim.ali@poliba.it) and Maria Pia Fanti (mariapia.fanti@poliba.it) are with the Department of Electrical and Information Engineering, Polytechnic University of Bari, 70125 Bari, Italy. Luis C. Félix-Herrán (lcfelix@tec.mx), is with the Tecnológico de Monterrey, Mexico.

in the preprocessing phase may have resulted in the loss of histopathologically relevant information, raising questions about the model's reliability in real world, unprocessed clinical settings. Simultaneously, radiomics have emerged as a powerful technique for extracting quantitative markers from medical images obtained via CT, MRI or ultrasound. Despite their potential, these advanced imaging techniques require expensive infrastructure, intricate segmentation protocols and high processing times, limiting their daily clinical applicability.

The aims of this paper are to address the issue of data scarcity by creating a unified dataset of 933 samples and to implement a hybrid imputation pipeline for handling missing values. Furthermore, this study compares several ML algorithms integrated into a two-level Stacking Ensemble architecture that employs Logistic Regression as the Meta Learner. This approach allowed the model to reach a perfect specificity (1.0000) through low-cost, easily accessible routine clinical and hematological parameters.

II. Materials and Methods

A. Dataset Description and Harmonization

The primary source for this study is the public dataset "Using Machine Learning to Predict Ovarian Cancer" (Mendeley Data V11), which can be found at https://data.mendeley.com/datasets/th7fztbrv9/11. This dataset originally comprises five distinct files (Supplementary Data 1-5) containing clinical and hematological parameters.

A rigorous pipeline (Fig. 1)was implemented to unify and harmonize all the raw sources that were available; meanwhile, the reference study focused on the analysis of a cohort of 349 patients. This process tripled the sample size, resulting in a final dataset of 933 samples, thereby effectively addressing the data scarcity problem common in medical literature.

The final dataset includes a binary target variable, denoted as `TYPE` — where 0 refers to benign ovarian tumors and 1 indicates malignant carcinomas — and 48 common independent features, which can be grouped into two different major categories:

- 19 hematological parameters: Derived from the complete blood count, including leukocyte indices (NEU), red blood cells and platelets;
- 29 biomarkers and parameters: These include demographic data (age, menopausal status), biochemical and metabolic indicators (CA-125, HE4) and liver function indicators.

Divergences in the handling of the Carbohydrate Antigen 72-4 (CA72-4) marker and the `SUBJECT_ID` variable (non-informative) were revealed in the course of the harmonization phase, specifically during an analytical comparison between the preprocessed training set (Supplementary Data 3) and the raw version (Supplementary Data 4). The exclusion of CA72-4 in the preprocessed file is due to its high percentage of missing values, the inclusion of which could have caused systematic bias. From a clinical standpoint, the low prevalence of the CA72-4 variant (despite being associated with mucinous ovarian cancer), compared to serous subtypes, suggests that the overall robustness of the classification system is not compromised if it's omitted [10].

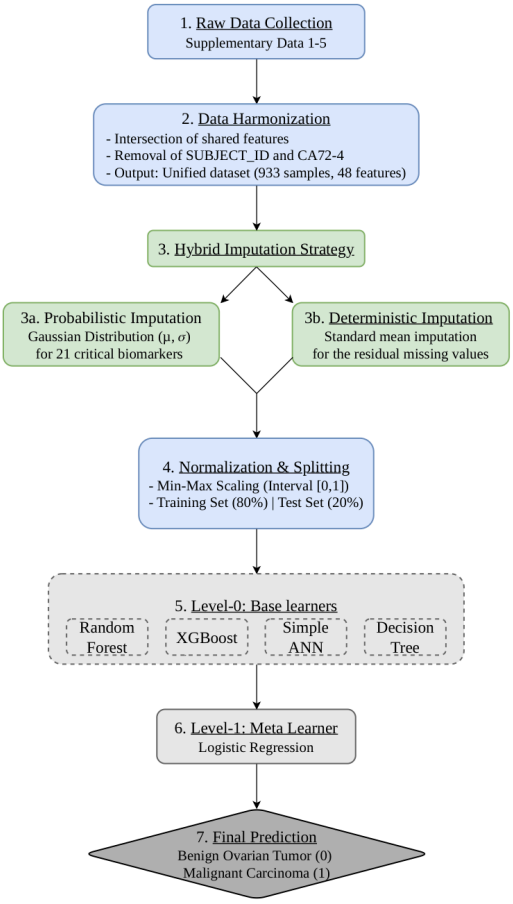

Fig. 1. Proposed methodology pipeline.

B. Data Preprocessing and Hybrid Imputation Strategy

A complex preprocessing phase was required in order to mitigate the structural heterogeneity of the initial data, which focused on the management of the missing values, and the normalization of the feature space [11]. The first step involved Feature Alignment; the Supplementary Data 2 file was excluded a priori due to its purely descriptive nature. Subsequently, an iterative intersection process was performed on the remaining four data frames, isolating 49 shared features. During this phase, the noninformative variables were removed; in this case indeed, the `SUBJECT_ID` variable, although useful for indexing, is useless when it comes to the predictive capacity of the models. Its inclusion would have induced rote learning (bias) in the model, causing overfitting and compromising its generalization capability on unseen patients. Consequently, the resulting feature set for analysis consists of 48 independent variables.

Due to the presence of many missing values, to preserve the variability of the clinical data, a hybrid imputation strategy

was adopted. The process is articulated into two hierarchical levels:

1) *Probabilistic Imputation Phase*: by using a stochastic approach (based on estimators derived from clinical reference range) missing values were filled for 21 critical biomarkers (e.g., CA-125, HE4, NEU). Specifically, each missing observation x_{ij} of feature j was assigned a value drawn from a Gaussian distribution:

$$x_{ij} \sim \max\big(0, \mathcal{N}(\mu_j, \sigma_j^2)\big) \tag{1}$$

where $\mathcal{N}$ denotes the Gaussian distribution with mean μ and variance σ^2. The application of a nonnegativity constraint ensures biological plausibility, while the use of predefined parameters (e.g., CA-125: $\mu = 17$, $\sigma = 9$ U/mL) prevents the artificial reduction of variance typically associated with standard mean imputation. This effectively combines clinically driven and stochastic data [12].

2) *Deterministic Imputation Phase*: To ensure the completeness of the dataset, standard mean imputation was applied; in this way the residual missing values can be handled [13]:

$$x_{ij} = \frac{1}{N} \sum_{k=1}^{N} x_{kj} \tag{2}$$

The dataset was split into a training set (80%) and a testing set (20%). Finally, to ensure that each feature contributes equally to the models' cost function, Min-Max Scaling normalization was applied:

$$x' = \frac{x - \min(x_{\text{train}})}{\max(x_{\text{train}}) - \min(x_{\text{train}})} \tag{3}$$

This rigorous approach — which consists in calculating the minimum and the maximum parameters exclusively on the training set and then applying them to the test set — allows preventing data leakage by scaling all values within the $[0, 1]$ interval; in this way, the convergence of ML algorithms is also improved.

All preprocessing steps, including imputation and scaling, were performed exclusively within the training folds during cross-validation to prevent data leakage. The test set remained completely unseen until final evaluation. In particular, imputation and scaling parameters were fitted separately within each training fold and applied to the corresponding validation fold.

C. Base Learners Description

Five heterogeneous ML algorithms — RF, SVM, DT, ANN, XGBoost — were implemented and optimized after dataset preprocessing and normalization. Scikit-learn library and XGBoost library were combined with a standard Python development environment to create the predictive models. For reproducibility purposes, the `random_state` parameter was set at 42 for all estimators, guaranteeing that the distribution between training and testing remains constant through every code execution [14]. Furthermore, given the need to maximise sensitivity due to the medical nature of the problem, the `class_weight = 'balanced'` parameter was set for the RF, SVM and DT models. This setting assigns a weight w_c to class c calculated as:

$$w_c = \frac{N}{k \cdot n_c} \tag{4}$$

where N is the total number of samples, k is the number of classes and n_c is the number of samples in class c [15]. The hyperparameters details are reported in Table I.

TABLE I
HYPERPARAMETERS OF THE BASE LEARNERS

Model	Optimized Hyperparameters
Random Forest	`n_estimators=80, max_depth=8, min_samples_split=5, min_samples_leaf=3, class_weight='balanced', random_state=42`
SVM	`kernel='rbf', C=0.5, gamma='scale', class_weight='balanced', probability=True, random_state=42`
Decision Tree	`max_depth=6, min_samples_split=8, min_samples_leaf=5, class_weight='balanced', random_state=42`
Simple ANN	`hidden_layer_sizes=(50,), activation='relu', solver='adam', alpha=0.001, learning_rate='adaptive', max_iter=500, random_state=42`
XGBoost	`n_estimators=80, max_depth=3, learning_rate=0.1, subsample=0.6, colsample_bytree=0.6, random_state=42`

The DT was selected for its capability to model nonlinear relationships through hierarchical partitions [16]. The quality of each split is determined by minimizing the impurity of the nodes via the Gini Index [17]:

$$\text{Gini Index}(t) = 1 - \sum_{i=1}^{c} [p(i \mid t)]^2 \tag{5}$$

where $p(i|t)$ represents the fraction of samples belonging to class i within node t. To overcome the limitation of a single tree, the RF aggregates the predictions of an ensemble of independent trees through bagging, drastically increasing the model's robustness [18].

The SVM algorithm aims to identify the optimal hyperplane that maximizes the separation margin between the classes in the normalized dataset, defined by the equation:

$$w^T x + b = 0 \tag{6}$$

where w is the weight vector, x is the input vector and b is the bias term. The use of Radial Basis Function (RBF) kernel allows the data to be mapped into a higher-dimensional space, facilitating the classification of complex, nonlinearly separable patterns [8].

XGBoost is a Gradient Boosting-based algorithm designed for its computational efficiency and speed. Unlike RF, XGBoost utilizes sequential boosting, where each new model,

$f_t(x)$, corrects the residual errors of the previous ones by optimizing a differentiable loss function:

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + \eta f_t(x_i) \tag{7}$$

where η is the learning rate and $\hat{y}_i^{(t)}$ is the prediction at iteration t [19].

Finally, the ANN was implemented, a multilayer neural network that optimizes synaptic weights to minimize classification error via backpropagation. To introduce nonlinearity and computational efficiency, the Rectified Linear Unit (ReLU) activation function was adopted in the hidden neurons:

$$\text{ReLU}(x) = \max(0, x) \tag{8}$$

which promotes convergence and mitigates the vanishing gradient problem [20].

D. Stacking Ensemble Architecture

To maximize the models' ability to generalize on new data and overcome the limitations of individual models, a Stacking Ensemble architecture was implemented. Unlike bagging or boosting techniques, Stacking combines independently trained heterogeneous models, called Base Learners, to create a more robust predictive Meta Learner [14]. The proposed architecture is structured into two hierarchical levels:

1) Level 0 (Base Learners): Based on the performance achieved during the validation phase (analyzed using the Gini Index and the Area Under the ROC Curve, AUC), the four best-performing models were selected: RF, XGBoost, ANN and DT. The SVM algorithm was excluded due to its inferior performance in terms of specificity. Each Base Learner M_j (with $j = 1, \ldots, 4$) receives the original feature vector x as input and returns a probability estimate $P_j(y = 1|x)$ for the malignant class. Utilizing probability distributions as meta-features allows the Meta Learner to exploit the predictive uncertainty of the underlying models to optimize the final decision.
2) Level 1 (Meta Learners): For the final aggregation phase, Logistic Regression was selected. The choice of a linear model enables learning the optimal weights to assign to each Base Learner, correcting potential biases of individual algorithms without introducing further risk of overfitting. The Meta Learner calculates the final probability $\hat{y}$ by linearly combining the prediction P_j:

$$\hat{y} = \sigma\left(\beta_0 + \sum_{j=1}^{4} \beta_j P_j\right) \tag{9}$$

where σ represents the sigmoid function, β_j represents the weight assigned to the prediction of the j-th model and β_0 is the intercept. A high β_j value indicates that Logistic Regression provides greater confidence to the predictions of the particular algorithm [21]. In addition, utilizing Logistic Regression as a Meta Learner ensures that the final Clinical Decision Support System (CDSS) is inherently well-calibrated. Unlike model types that only provide rigid and fixed labels, this architecture produces continuous and uniform probability distributions, as essential requirement in medical diagnostic system to correctly quantify predictive uncertainty.

To prevent data leakage and ensure reliable evaluations, a Stratified 5-fold Cross-Validation procedure was implemented. The training dataset was partitioned into 5 folds while maintaining the original class proportions: iteratively, the Base Learners are trained on 4 folds and generate predictions on the remaining fold. This ensures that the predictions used to train Logistic Regression are strictly out-of-fold. This approach guarantees that the Meta Learner operates on effectively out-of-sample data, preventing overfitting and ensuring the statistical validity of the ensemble [22], [23].

E. Evaluation Metrics

To objectively evaluate and compare the performance of individual classifiers and the Stacking Ensemble architecture, several standard metrics derived from the Confusion Matrix were employed. The quantitative analysis is based on four fundamental parameters: True Positives (TP), True Negatives (TN), False Positives (FP) and False Negatives (FN).

The implemented metrics are defined as follows [25]:

1) Accuracy: Evaluates the percentage of correct prediction relative to the total number of analyzed samples.

$$\text{Accuracy} = \frac{TP + TN}{TP + FP + TN + FN} \tag{10}$$

2) Sensitivity (Recall): Quantifies the model's ability to correctly identify all malignant cases ($TYPE = 1$). In oncology, maximizing this metric is a priority to minimize FN, preventing fatal delays in life-saving treatments.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \tag{11}$$

3) Specificity: Evaluates the algorithm's ability to recognize benign cases ($TYPE = 0$). High specificity is essential to limit FP, which helps avoid severe psychological stress and unnecessary invasive diagnostic procedures.

$$\text{Specificity} = \frac{TN}{TN + FP} \tag{12}$$

4) F1-Score: Represents the harmonic mean between Sensitivity and Precision ($TP/(TP + FP)$), providing a balanced assessment of the model's reliability.

$$\text{F1-Score} = 2 * \frac{(Precision * Sensitivity)}{Precision + Sensitivity} \tag{13}$$

Beyond point metrics, the overall discriminatory capacity is quantified by the Receiver Operating Characteristic (ROC) curve and its corresponding AUC. The ROC curve plots Sensitivity against the FP rate ($1 - Specificity$) across various confidence thresholds. An AUC value close to 1.0000 indicates a classifier with excellent discriminatory capabilities. Closely related to the AUC is the Gini Index, a statistical metric used to quantify the direct discriminatory power of the classification

model. Finally, to mitigate the "black-box" nature of complex models, Feature Importance analysis was conducted to extract the decision weight of the most relevant clinical and hematological biomarkers.

III. Results and Discussion

A. Performance Analysis and Model Comparison

The performance of the predictive models was evaluated on the test set (comprising of 20% of the original data), comparing the individual Base Learners with the Stacking Ensemble architecture across various classification metrics. As summarized in Table II, the Stacking approach outperformed all classification models across all considered parameters.

TABLE II
PERFORMANCE COMPARISON (FROM BEST TO WORST) OF INDIVIDUAL MODELS VS STACKING ENSEMBLE.

Method	Accuracy	Sensitivity	Specificity	F1-Score	AUC	Gini
Stacking	**0.9840**	**0.9663**	**1.0000**	**0.9829**	**0.9990**	**0.9979**
XGBoost	0.9733	0.9663	0.9796	0.9718	0.9989	0.9977
Random Forest	0.9786	0.9663	0.9898	0.9773	0.9986	0.9972
Simple ANN	0.9519	0.9663	0.9388	0.9503	0.9913	0.9826
Decision Tree	0.9572	0.9551	0.9592	0.9551	0.9764	0.9529
SVM	0.9037	0.9551	0.8571	0.9043	0.9701	0.9402

While robust algorithms such as XGBoost or RF demonstrated excellent discriminatory performance (achieving a Gini Index of 0.9977 and 0.9972, respectively), the integration of the Stacking method allowed for further refinement of the decision boundary, reaching a Gini Index of 0.9979. By learning the strengths and limitations of the individual underlying models — such as the low specificity observed in the SVM (0.8571) — the Meta Learner achieved an overall accuracy of 0.9840.

B. Error Minimization: Confusion Matrix and ROC Curve Analysis

From a clinical perspective, analyzing error through the Confusion Matrix highlights the added value of the proposed architecture compared to individual classifiers. In oncology, it is crucial that a CDSS minimize FN in order to avoid fatal delays in life-saving treatments. As evidenced by the experimental results, the best-performing models (Stacking, XGBoost, RF, ANN) maintained a high sensitivity of 0.9663, correctly identifying 86 out of 89 malignant cases in the test set. However, the most significant achievement lies in the total elimination of FP, achieved exclusively by the Stacking model, which attained a perfect specificity of 1.0000. This result is of fundamental importance, as correctly classifying benign cases means avoiding psychological stress for the patient and preventing unnecessary, invasive diagnostic procedures. This discriminatory ability is confirmed by the analysis of the ROC curve (Fig. 2), whose profile closely approaches the top-left corner of the graph, reaching an AUC of 0.9990.

Although the standard results are reported using a 0.5 classification threshold, the well-calibrated probabilities generated by the Meta Learner allow for a dynamic adjustment of the threshold itself. The system enables clinicians to deliberately lower this threshold to maximize Sensitivity in high-risk contexts, ensuring flexible and safe management of the clinical trade-off between early diagnosis and diagnostic precision.

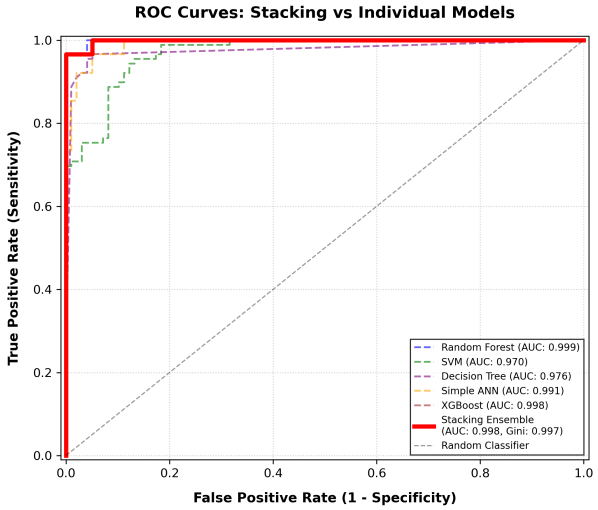

Fig. 2. ROC curve of the Stacking Ensemble model (AUC = 0.9990)

C. Feature Importance and Biomarker Relevance

The interpretability of tree-based models (RF, DT, XGBoost) allowed us to extract the decision weight of individual features, validating ML against established medical literature. The Feature Importance analysis reveals strong consistency among the different algorithms, with some differences that highlight the uniqueness of each approach. The dominant predictors were found to be NEU, age and menopausal status (Fig. 3).

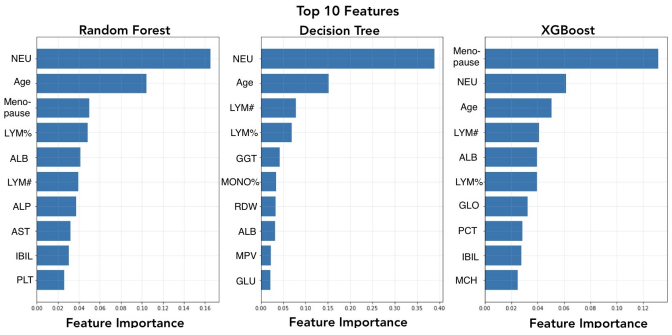

Fig. 3. Feature Importance ranking extracted from the tree-based models.

In the DT model, the NEU feature constitutes the key element for the initial data partitioning, demonstrating immediate discriminatory power. Often analyzed in conjunction with NEU, alterations in leukocyte indices — particularly the lymphocyte ratio (with LYM% also being among the most relevant parameters) — are well known systemic indicators of the inflammatory response associated with neoplastic proliferation and tumor immunosuppression. The XGBoost algorithm identified menopausal status as the primary factor. This

underscores the critical importance of a patient's hormonal and demographic status in the development of ovarian cancer. Furthermore, the importance associated with biochemical parameters such as albumin and alkaline phosphatase further confirms the ability of AI to capture complex metabolic and liver function alterations. Finally, the integration of these routine hematological and systemic parameters with classic tumor biomarkers (such as CA-125, HE4) through the Stacking Ensemble architecture allows for a more refined decision-making process, maximizing diagnostic accuracy, especially in clinical cases characterized by nonspecific symptomatology.

D. Clinical Implications and Future Perspectives

The conducted analysis has demonstrated the effectiveness of AI in extracting high-precision diagnostic patterns by analyzing routine clinical and hematological parameters. The implementation of the Stacking Ensemble architecture represents a concrete response to the growing need for rapid diagnostic tools in contexts characterized by a shortage of medical specialists and pathologists. In resource-limited environments, a CDSS utilizing this model can act as a highly reliable diagnostic tool, accelerating diagnostic times and reducing human error related to subjective interpretation. However, it is important to emphasize that the proposed model does not aim to replace the physician's judgment, especially in complex and dynamic contexts; rather, it seeks to provide specialists with a quantitative second opinion based on objective probabilistic estimates. The synergy between the computational power of Ensemble algorithms and the physician's clinical experience represents the most promising prospect for integrating AI into daily hospital practice, paving the way for increasingly earlier and more personalized diagnoses [24].

IV. Conclusion

This study demonstrated the effectiveness of AI, with a particular focus on advanced Ensemble Learning techniques, as a diagnostic support tool for ovarian cancer. An essential contribution of this work lies in overcoming data scarcity limitations frequently encountered in recent literature. By harmonizing routine clinical and hematological data and implementing a hybrid preprocessing pipeline, a robust predictive framework was established, capable of addressing the heterogeneity and missing values typical of clinical settings. The comparative analysis highlighted how the hierarchical Stacking Ensemble architecture represents a significant performance leap compared to the state of the art. By combining the probabilistic predictions of the best individual classifiers (XGBoost, RF, ANN and DT) through a Logistic Regression-based Meta Learner, the system achieved exceptional results, reaching an accuracy of 0.9840, an AUC of 0.9990 and a Gini Index of 0.9979. From a clinical perspective, the most significant Stacking result is the achievement of perfect specificity (1.0000), which involves the complete elimination of FPs while maintaining a high sensitivity of 0.9663. This demonstrates the model's ability to optimize the decision boundary, avoiding unnecessary invasive procedures for healthy patients, while optimizing the detection of malignant cases using cost-effective biomarkers (NEU, CA-125, HE4) and demographic information [8].

References

[1] S. Mitchell *et al.*, "Artificial intelligence in ultrasound diagnoses of ovarian cancer: a systematic review and meta-analysis," *Cancers*, vol. 16, no. 2, p. 422, 2024.

[2] Z. Momenimovahed, A. Tiznobaik, S. Taheri, and H. Salehiniya, "Ovarian cancer in the world: epidemiology and risk factors," *World Journal of Oncology*, vol. 10, no. 1, pp. 8–15, 2019.

[3] D. M. Robertson, "Screening for the early detection of ovarian cancer," *Women's Health*, vol. 5, pp. 347–349, 2009.

[4] W.-H. Wang *et al.*, "Systematic review and meta-analysis of imaging differential diagnosis of benign and malignant ovarian tumors," *Eur. Rev. Med. Pharmacol. Sci.*, vol. 26, no. 12, pp. 4181–4191, 2022.

[5] V. Dambra, M. Roccotelli and M. P. Fanti, "Diabetic Disease Detection using Machine Learning Techniques," 2024 10th International Conference on Control, Decision and Information Technologies (CoDIT), Vallette, Malta, 2024, pp. 1436-1441.

[6] I. P. Battista, M. Roccotelli, W. Ali and M. P. Fanti, "Diagnosis of Parkinson's Disease Using Machine Learning Algorithms," 2025 11th International Conference on Control, Decision and Information Technologies (CoDIT), Split, Croatia, 2025, pp. 200-205.

[7] M. Akazawa and K. Hashimoto, "Artificial intelligence in ovarian cancer diagnosis," *Anticancer Research*, vol. 40, no. 8, pp. 4795–4800, 2020.

[8] S. M. Ayyoubzadeh, M. Ahmadi, A. B. Yazdipour, F. Ghorbani-Bidkorpeh, and M. Ahmadi, "Prediction of ovarian cancer using artificial intelligence tools," *Health Sci. Rep.*, vol. 7, no. 7, Art. no. e2203, 2024.

[9] P. Paayas and R. Annamalai, "Ocean - ovarian cancer subtype classification and outlier detection using densenet121," 2023, pp. 827–831.

[10] U. Matulonis, A. Sood, L. Fallowfield, *et al.*, "Ovarian cancer," *Nat. Rev. Dis. Primers*, vol. 2, p. 16061, 2016, doi: 10.1038/nrdp.2016.61.

[11] S. K. Kwak and J. H. Kim, "Statistical data preparation: management of missing values and outliers," *Korean J. Anesthesiol.*, vol. 70, no. 4, pp. 407–411, Aug. 2017.

[12] A. Waljee *et al.*, "Comparison of imputation methods for missing laboratory data in medicine," *BMJ Open*, vol. 3, no. 8, Aug. 2013.

[13] R. J. A. Little and D. B. Rubin, *Statistical Analysis With Missing Data*, 3rd ed. Hoboken, NJ, USA: John Wiley & Sons, 2019.

[14] F. Pedregosa *et al.*, "Scikit-learn: Machine learning in Python," *J. Mach. Learn. Res.*, vol. 12, pp. 2825–2830, 2011.

[15] H. Liu, M. Zhou, X. S. Lu, and C. Yao, "Weighted Gini index feature selection method for imbalanced data," in *2018 IEEE 15th Int. Conf. Netw., Sens. Control (ICNSC)*, 2018, pp. 1–6.

[16] S. Singh and P. Gupta, "Comparative study ID3, CART and C4.5 decision tree algorithm: A survey," *Int. J. Adv. Inf. Sci. Technol. (IJAIST)*, vol. 27, no. 27, pp. 97–103, 2014.

[17] I. D. Mienye and N. Jere, "A survey of decision trees: Concepts, algorithms, and applications," *IEEE Access*, vol. 12, pp. 86716–86727, 2024.

[18] X. Chen and H. Ishwaran, "Random forests for genomic data analysis," *Genomics*, vol. 99, no. 6, pp. 323–329, 2012.

[19] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 785–794.

[20] S. Walczak, "Artificial neural networks," in *Adv. Methodol. and Technol. in Artif. Intell., Comput. Simulation, and Human-Comput. Interact.* Hershey, PA, USA: IGI Global, 2019, pp. 40–53.

[21] M. A. Ganaie, M. Hu, A. K. Malik, M. Tanveer, and P. N. Suganthan, "Ensemble deep learning: A review," *Eng. Appl. Artif. Intell.*, vol. 115, Art. no. 105151, 2022.

[22] S. Q. Sultan, N. Javaid, N. Alrajeh, and M. Aslam, "Machine learning-based stacking ensemble model for prediction of heart disease with explainable AI and k-fold cross-validation: A symmetric approach," *Symmetry*, vol. 17, no. 2, p. 185, 2025.

[23] T. Leinonen *et al.*, "Empirical investigation of multi-source cross-validation in clinical ECG classification," *Comput. Biol. Med.*, vol. 183, Art. no. 109271, 2024.

[24] P. Rajpurkar *et al.*, "AI in health and medicine," *Nat. Med.*, vol. 28, no. 1, pp. 31–38, 2022.

[25] M. Hossin and M. N. Sulaiman, "A review on evaluation metrics for data classification evaluations," *Int. J. Data Mining & Knowl. Manage. Process*, vol. 5, no. 2, p. 1, 2015.