Enhanced Sepsis Prediction Using Ensemble Learning with SMOTE-Based Data Balancing and Stratified Validation

N. Smitha
Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, India | Bangalore University, Bengaluru, India
smithan.ckm@gmail.com (corresponding author)

R. Tanuja
Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, India | Bangalore University, Bengaluru, India
tanujar.uvce@gmail.com

S. H. Manjula
Department of Computer Science and Engineering, University Visvesvaraya College of Engineering, India | Bangalore University, Bengaluru, India
shmanjula@gmail.com

Received: 14 August 2025 | Revised: 14 October 2025 and 28 October 2025 | Accepted: 31 October 2025

 DOI: https://doi.org/10.48084/etasr.14071

ABSTRACT

Sepsis, a critical condition triggered by an abnormal immune response to infection, requires rapid and accurate identification to reduce the risk of mortality. In clinical settings, datasets often suffer from severe class imbalance, with sepsis cases significantly underrepresented, which complicates early prediction efforts. This study explores and compares the effectiveness of traditional and ensemble-based machine learning algorithms for sepsis detection. Initially, models such as Random Forest (RF), Support Vector Machine (SVM), XGBoost, and K-Nearest Neighbors (KNN) were trained on imbalanced datasets using median imputation. To improve model reliability and address class imbalance, advanced ensemble techniques, such as soft-voting and stacking, were incorporated, along with cost-sensitive learning and stratified k-fold validation. The Synthetic Minority Oversampling Technique (SMOTE) was later applied to balance the dataset, and the models were reassessed. Evaluation based on metrics such as ROC-AUC, PR-AUC, sensitivity, specificity, balanced accuracy, and Brier score revealed that the stacking ensemble, applied to SMOTE-processed data, delivered superior performance with a ROC-AUC of 0.9979. These results show how ensemble approaches and data balancing strategies can improve the precision and dependability of sepsis prediction models used in clinical decision-making.

Keywords-ensemble learning; sepsis detection; stacking; SMOTE; ROC; AUC

I. INTRODUCTION

Organ dysfunction in sepsis, a potentially lethal disease, is caused by a deregulated host response to infection. Detecting sepsis in a timely and accurate manner is of paramount importance, as early intervention significantly reduces mortality and long-term complications [1]. However, detecting sepsis remains challenging due to its heterogeneous clinical presentation, evolving physiological markers, and the imbalanced nature of real-world clinical datasets, where instances of sepsis are significantly outnumbered by non-sepsis ones.

ML has emerged as a valuable decision-support tool in the early detection of sepsis, offering data-driven approaches to predict its onset based on patient vitals, laboratory results, and other clinical features. Traditional Machine Learning (ML) models, such as Random Forest (RF) [2], Support Vector Machine (SVM) [3], and Extreme Gradient Boosting (XGBoost), have demonstrated considerable promise [4]. Conventional ML classifiers generally favor the majority class, showing low sensitivity and poor recall in the minority class. In addition, when applied separately, ML models often face problems such as bias or high variance, restricting their ability to generalize across a range of clinical conditions and patient populations.

To address these limitations, ensemble learning techniques have been proposed, in which a more reliable predictive model is produced by combining many base learners [5]. Among the various ensemble strategies, soft voting (probabilistic averaging of multiple classifiers) [6] and stacking (meta-learning over base model predictions) have gained interest. These methods can leverage the strengths of multiple models and produce a more sophisticated and accurate prediction pipeline by balancing the drawbacks of each model.

In addition, dealing with data imbalance is critical to training effective classifiers. Synthetic Minority Oversampling Technique (SMOTE) has proven effective in creating a balanced training distribution by generating synthetic examples for the minority class, thereby enhancing model sensitivity without sacrificing specificity [7]. However, the effectiveness of ensemble models trained on SMOTE-resampled data—especially in high-stakes healthcare applications—has not been systematically evaluated.

This work presents a comparative analysis of individual classifiers (RF, SVM, XGBoost) and ensemble models (soft-voting and stacking) for the early detection of sepsis. All models are trained and evaluated on clinical data balanced using SMOTE. This allows assessing not only discriminative ability but also calibration and class-wise performance. The study's conclusions show that ensemble strategies—especially stacking—consistently perform better than single-model techniques, especially in capturing minority class instances (sepsis cases) while maintaining calibration and low false-positive rates.

Existing research on early sepsis prediction using ML demonstrates promising results but reveals several gaps in methodological integration and evaluation. Previous studies often focused on either data balancing through SMOTE or model enhancement via ensemble learning, but lacked a unified framework to ensure both robust validation and cost-sensitive learning. Additionally, most works relied on limited metrics, such as accuracy and AUC, neglecting critical clinical measures such as sensitivity, specificity, PR AUC, and Brier score, and rarely applied stratified k-fold validation to ensure generalizability across heterogeneous patient populations. In addition to addressing these drawbacks, the novelty of this study lies in its thorough and unified approach that integrates SMOTE-based data balancing, cost-sensitive ensemble learning (voting and stacking), and stratified validation, forming a robust pipeline for reliable sepsis detection.

II. METHODOLOGY

This section presents the detailed method adopted to develop an enhanced sepsis prediction framework that integrates SMOTE-based data balancing, ensemble learning, and stratified cross-validation. The proposed model aims to achieve early and accurate detection of sepsis by addressing issues such as data imbalance, model bias, and poor generalization observed in earlier works, as shown in Figure 1.

A. Dataset Preparation

The dataset, derived from the PhysioNet Sepsis Challenge [8], includes vital signs and lab results. The data was preprocessed using median imputation and feature normalization. The PhysioNet Sepsis dataset was introduced in the 2019 PhysioNet/Computing in Cardiology Challenge, titled "Early Prediction of Sepsis from Clinical Data." Each row represents one hour of a patient's ICU stay and includes 40 clinical variables along with an outcome label. The features span across demographics (such as age, gender, and ICU type), vital signs (heart rate, oxygen saturation, temperature, blood pressures, and respiratory rate), laboratory measurements (such as glucose, lactate, creatinine, BUN, WBC, and platelets), and other physiological parameters, including electrolytes and gas exchange variables.

B. Data Preprocessing

Since the raw data contained missing and noisy values, multiple preprocessing steps were implemented:

- Missing value imputation: Continuous variables were filled using median values or forward-fill interpolation and categorical variables were imputed with mode [9].
- Outlier handling: Statistical thresholds (IQR and Z-score) were used to smooth extreme outliers.
- Normalization: All continuous features were scaled using min-max normalization to ensure uniformity for model training.
- Feature selection: Highly correlated and redundant features were eliminated using correlation analysis and importance ranking through RF.

TABLE I. REVIEW OF PREVIOUS WORKS

Study	Year	Study title/focus	Dataset(s)	Techniques used	Best model /approach	Performance metrics	Key contributions
[10]	2024	Early sepsis prediction with unbalanced data processing	2,385 patients (First Affiliated Hospital of Anhui Medical University) + MIMIC-III + eICU	Multiple imputation, SMOTE, SHAP, ML classifiers	Random Forest	AUC = 0.87, F1 = 0.77	Addressed data imbalance with SMOTE, identified key features (SBP, Albumin, HR), validated across multiple datasets.
[11]	2025	Machine Learning	PhysioNet 2019	Data preprocessing, feature extraction, ML classifier training	ML-SVM	Accuracy = 95.2%, Sensitivity = 91%, Specificity = 93%	Proposed an ML-SVM model to predict sepsis onset using ICU records, demonstrated reliable early detection.
[12]	2025	ML-based early detection of neonatal sepsis	Neonatal clinical records	ML classifiers	Ensemble learning	Accuracy = 98.6%, Precision = 97%	Improved diagnostic accuracy and reliability by integrating multiple classifiers.

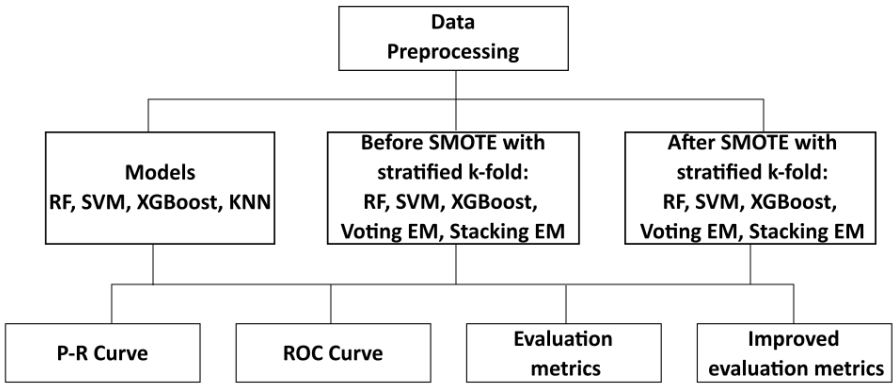

Fig. 1. Workflow diagram of sepsis prediction pipeline using ensemble learning before and after SMOTE with stratified k-fold validation.

C. Data Balancing Using SMOTE

The dataset was highly imbalanced, with sepsis-positive cases representing less than 10% of the total records. SMOTE was applied to prevent model bias toward the majority class. SMOTE synthetically generates new minority class instances based on the feature-space similarities between existing samples, thereby improving the classifier's ability to identify rare sepsis cases. Mathematically, new synthetic samples are generated as:

$$x_{\{new\}} = x_i + \delta \times (x_{\{nn\}} - x_i) \quad (1)$$

where x_i is a minority sample, $x_{\{nn\}}$ is one of its k nearest neighbors, and $\delta \in [0,1]$.

D. Baseline Model Evaluation on Imputed Data

At first, four popular ML models were evaluated on the imputed dataset without using any resampling or cost-sensitive modifications:

- Random Forest (RF): A tree-based ensemble model that aggregates predictions from several decision trees trained on random feature subsets to reduce overfitting [13].
- K-Nearest Neighbors (KNN): A non-parametric model that classifies based on the majority class among the k-nearest data points in the feature space [14].
- Support Vector Machine (SVM): For high-dimensional data, a margin-based classifier works well. For non-linear separation, a Radial Basis Function (RBF) kernel was employed [15].
- XGBoost (Extreme Gradient Boosting): A highly efficient boosting algorithm that sequentially builds decision trees while minimizing a loss function with regularization [16].

E. Cost-Sensitive Learning Before SMOTE

- To address class imbalance (i.e., fewer sepsis cases), cost-sensitive learning was incorporated into the models before applying oversampling. During training with this method, misclassification of minority class samples is penalized more severely.
- For SVM, $class\_weight = 'balanced'$ was used to scale the penalty parameter C differently for each class.
- In XGBoost, $scale\_pos\_weight$ was set to the ratio of negative to positive samples, boosting focus on the minority class.

These steps improve sensitivity (true positive rate) without the risk of overfitting that may come from oversampling.

F. Ensemble Learning: Voting and Stacking

To further enhance performance, two ensemble learning strategies were implemented:

- Soft-voting combines the predicted probabilities of multiple base models (RF, SVM, XGB) and chooses the class with the greatest average likelihood. It improves calibration and reduces variance by leveraging diverse models.
- Stacking uses a meta-model (typically logistic regression) to learn from the output probabilities of base models. This method captures higher-order interactions among model outputs, improving generalization performance as shown in:

  $$z = f_{meta}(h_1(x), h_2(x), \ldots, h_n(x)) \quad (2)$$

  where $h_1(x), h_2(x), \ldots, h_n(x)$ are the outputs of the n individual base models, each trained independently on the same input x, and z is the final output or class prediction made by the stacking ensemble.

Both ensembles are trained with and without resampling to assess their robustness under different data distributions.

G. Stratified K-Fold Cross-Validation

This method was selected to maintain class distribution in every fold and guarantee accurate performance estimation. Each of the $k = 5$ folds of the data maintains the same percentage of sepsis cases compared to non-sepsis cases [17]. This method reduces the possibility of biased evaluation that could happen when using simple random splits in imbalanced datasets.

III. RESULTS AND DISCUSSION

A. Experimental Setup Abbreviations and Acronyms

This study focuses on enhanced sepsis detection using ensemble learning techniques applied to a real-world clinical dataset. Using both balanced and unbalanced data, the experimental process aimed to evaluate the performance of

base and ensemble models. Performance analysis evaluates how well ensemble and conventional ML models detect sepsis under three experimental settings:

- Baseline, with the original imbalanced data.
- Before SMOTE, with stratified k-fold and cost-sensitive learning.
- After SMOTE, with stratified k-fold resampling.

In the baseline setup, models were trained on imbalanced clinical data, where sepsis cases were significantly underrepresented. Initially, this setup appeared to perform well, with high accuracy and recall across all four classifiers (KNN, SVM, RF, and XGBoost). However, these metrics were misleading due to the dataset imbalance, which caused the models to favor the majority (non-sepsis) class. This issue was clearly reflected in the low ROC-AUC scores, particularly for KNN (0.5351), indicating poor discriminative ability. Although RF and SVM achieved slightly better AUCs (~0.70), they still struggled to generalize well to the minority class. The results in Table II highlight the limitations of relying solely on conventional metrics in imbalanced settings and demonstrate the need for resampling techniques, such as SMOTE, and advanced ensemble methods to improve true positive detection and model calibration.

TABLE II. BASELINE MODEL PERFORMANCE (IMBALANCED DATA, MEDIAN IMPUTATION)

Models	Accuracy	Precision	Recall	F1-score	ROC-AUC
KNN	0.93	0.86	0.93	0.89	0.53
SVM	0.93	0.86	0.93	0.89	0.69
RF	0.92	0.86	0.92	0.89	0.70
XGBoost	0.93	0.92	0.93	0.91	0.59

Although the cost-sensitive classifiers improved ROC-AUC marginally over the baseline in some cases, the Precision-Recall (P-R) tradeoff remained poor. Many models failed to identify positive sepsis cases (recall = 0), indicating persistent skewness in model behavior toward the majority class, as shown in Table III.

TABLE III. BEFORE SMOTE WITH STRATIFIED K-FOLD (COST-SENSITIVE MODELS)

Model	Accuracy	Precision	Recall	F1-score	ROC-AUC
KNN	0.93	0.00	0.00	0.00	0.54
RF	0.93	0.00	0.00	0.00	0.93
SVM	0.61	0.08	0.47	0.14	0.49
XGB	0.92	0.18	0.02	0.04	0.69
VotingEM	0.93	0.00	0.00	0.00	0.66
Stacking EM	0.93	0.00	0.00	0.00	0.63

The results in Table IV show considerable improvement in model performance across all measures. Notably, the recall and ROC AUC scores surged for most classifiers, confirming the positive impact of addressing class imbalance. XGBoost and RF, in particular, showed substantial improvements, with XGBoost achieving an accuracy of 97.20% and a ROC-AUC of 0.9966. The ensemble models further elevated performance, with the soft voting ensemble achieving a balanced F1-score of 0.9714 and a ROC-AUC of 0.9950.

TABLE IV. AFTER SMOTE WITH STRATIFIED K-FOLD (BALANCED DATA)

Models	Accuracy	Precision	Recall	F1-score	ROC AUC
KNN	0.96	0.96	0.96	0.96	0.96
SVM	0.58	0.58	0.58	0.58	0.58
RF	0.97	0.97	0.97	0.97	0.97
XGB	0.97	0.97	0.97	0.97	0.97
Stacking EM	0.97	0.97	0.96	0.97	0.99
Voting EM	0.97	0.97	0.97	0.97	0.99

Most impressively, the stacking ensemble outperformed all other models, recording the highest accuracy (97.80%), F1-score (0.9780), and ROC-AUC (0.9979), indicating exceptional discriminative ability and balanced classification as shown in Table IV and Figures 2, 3, and 4. In the PR curve, RF, XGBoost, and the voting and stacking ensembles demonstrate nearly perfect performance, with AUC values close to 1.00, indicating strong predictive precision even at high recall levels. The SVM, however, shows much lower PR-AUC (0.63), revealing its limitations in correctly identifying the minority (sepsis) class.

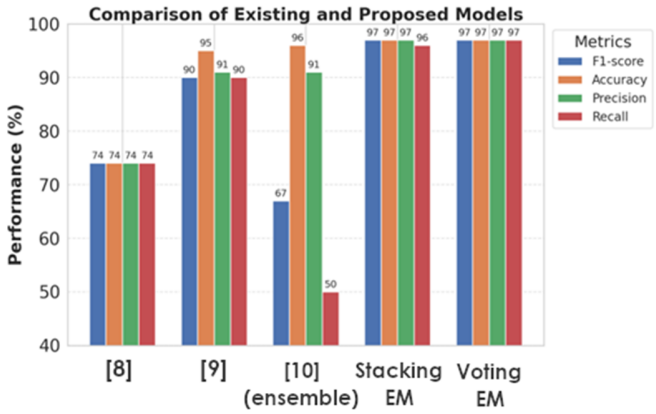

Fig. 2. Comparison of existing and proposed models.

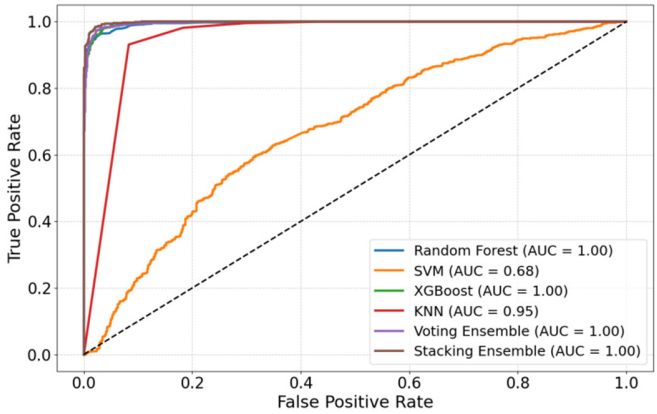

Fig. 3. ROC curves for all models.

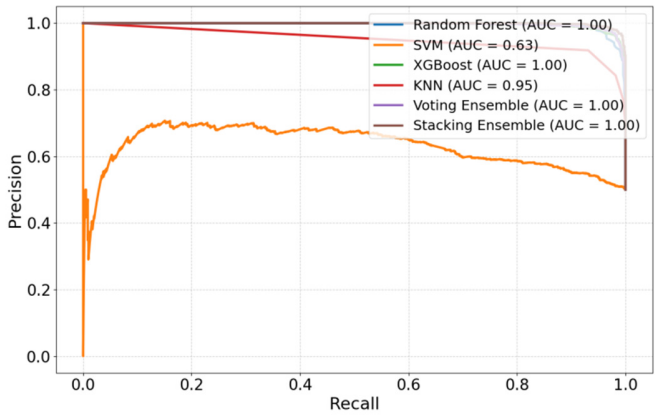

Fig. 4. Precision-recall curves for all models.

TABLE V. RESULTS FOR SEPSIS DETECTION MODELS ON IMBALANCED CLINICAL DATA USING STRATIFIED K-FOLD CROSS-VALIDATION (WITHOUT SMOTE).

Model	ROC-AUC	PR-AUC	Sensitivity	Specificity	Balanced accuracy	Brier score
KNN	0.94	0.95	0.93	0.91	0.92	0.11
RF	0.99	0.99	0.96	0.97	0.96	0.05
SVM	0.67	0.62	0.93	0.23	0.58	0.22
XGB	0.99	0.99	0.98	0.96	0.97	0.02
Voting EM	0.99	0.99	0.96	0.97	0.97	0.06
Stacking EM	0.99	0.99	0.97	0.97	0.97	0.01

IV. LIMITATIONS AND FUTURE WORK

This work has certain limitations, although ensemble learning models in conjunction with SMOTE-based balancing show promising results. Initially, only one dataset was used to train and assess the models, which can restrict their applicability in other medical settings. Additionally, the SMOTE technique, while effective in addressing class imbalance, can introduce synthetic noise and does not preserve the temporal patterns inherent in sequential ICU data. These models also treat each patient snapshot independently, lacking the temporal context crucial in progressive conditions such as sepsis. Future research will focus on incorporating time-series modeling approaches such as LSTM and GRU. More advanced sampling strategies, such as SMOTE-ENN or ADASYN, could be explored to minimize overfitting from synthetic samples [18].

V. CONCLUSION

This study presents a comprehensive framework that combines SMOTE-based data balancing, cost-sensitive ensemble learning, and stratified k-fold cross-validation, ensuring fair representation of sepsis and non-sepsis cases and improving model robustness. Multiple classifiers, including RF, SVM, XGBoost, and KNN, were trained and combined using ensemble methods such as voting and stacking. After applying SMOTE and stratified validation, the ensemble models—especially the stacking approach—achieved excellent performance, with ROC-AUC and PR-AUC of 0.99, F1-score of 0.97, sensitivity and specificity of 0.97, balanced accuracy of 0.97, and Brier score of 0.01, highlighting accurate early detection while minimizing false predictions.

REFERENCES

[1] D. B. Gotur, "Sepsis Diagnosis and Management," *Journal of Medical Sciences and Health*, vol. 03, no. 03, pp. 1–12, Dec. 2017, https://doi.org/10.46347/JMSH.2017.v03i03.001.

[2] S. J. Rigatti, "Random Forest," *Journal of Insurance Medicine*, vol. 47, no. 1, pp. 31–39, 2017, https://doi.org/10.17849/insm-47-01-31-39.1.

[3] H. Xue, Q. Yang, and S. Chen, "SVM: Support Vector Machines," in *The Top Ten Algorithms in Data Mining*, Chapman and Hall/CRC, 2009.

[4] T. Chen *et al.*, "xgboost: Extreme Gradient Boosting." May 15, 2025, [Online]. Available: https://cran.r-project.org/web/packages/xgboost/.

[5] G. I. Webb and Z. Zheng, "Multistrategy ensemble learning: reducing error by combining ensemble learning techniques," *IEEE Transactions on Knowledge and Data Engineering*, vol. 16, no. 8, pp. 980–991, Aug. 2004, https://doi.org/10.1109/TKDE.2004.29.

[6] S. Kumari, D. Kumar, and M. Mittal, "An ensemble approach for classification and prediction of diabetes mellitus using soft voting classifier," *International Journal of Cognitive Computing in Engineering*, vol. 2, pp. 40–46, June 2021, https://doi.org/10.1016/j.ijcce.2021.01.001.

[7] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "SMOTE: Synthetic Minority Over-sampling Technique," *Journal of Artificial Intelligence Research*, vol. 16, pp. 321–357, June 2002, https://doi.org/10.1613/jair.953.

[8] M. Reyna *et al.*, "Early Prediction of Sepsis from Clinical Data: The PhysioNet/Computing in Cardiology Challenge 2019." PhysioNet, https://doi.org/10.13026/V64V-D857.

[9] S. Tripathi, L. Singh, and J. Sermanraja, "Complete Data Using Exploratory Data Analysis and ML Algorithms," in *2024 4th International Conference on Technological Advancements in Computational Sciences (ICTACS)*, Tashkent, Uzbekistan, Nov. 2024, pp. 1293–1296, https://doi.org/10.1109/ICTACS62700.2024.10840870.

[10] L. Zhou, M. Shao, C. Wang, and Y. Wang, "An early sepsis prediction model utilizing machine learning and unbalanced data processing in a clinical context," *Preventive Medicine Reports*, vol. 45, Sept. 2024, Art. no. 102841, https://doi.org/10.1016/j.pmedr.2024.102841.

[11] R. M. A. El-Aziz and A. Rayan, "Early detection of sepsis using machine learning algorithms," *Alexandria Engineering Journal*, vol. 111, pp. 47–56, Jan. 2025, https://doi.org/10.1016/j.aej.2024.10.005.

[12] S. A. Parvin and B. Saleena, "Designing a hybrid stack ensemble model to enhance sepsis classification using data triangulation approach," *Results in Engineering*, vol. 25, Mar. 2025, Art. no. 103748, https://doi.org/10.1016/j.rineng.2024.103748.

[13] M. Belgiu and L. Drăguţ, "Random forest in remote sensing: A review of applications and future directions," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 114, pp. 24–31, Apr. 2016, https://doi.org/10.1016/j.isprsjprs.2016.01.011.

[14] L. Peterson, "K-nearest neighbor," *Scholarpedia*, vol. 4, no. 2, 2009, Art. no. 1883, https://doi.org/10.4249/scholarpedia.1883.

[15] W. S. Noble, "What is a support vector machine?," *Nature Biotechnology*, vol. 24, no. 12, pp. 1565–1567, Dec. 2006, https://doi.org/10.1038/nbt1206-1565.

[16] T. Chen *et al.*, "xgboost: Extreme Gradient Boosting." Sept. 01, 2014, https://doi.org/10.32614/CRAN.package.xgboost.

[17] X. Zeng and T. R. Martinez, "Distribution-balanced stratified cross-validation for accuracy estimation," *Journal of Experimental & Theoretical Artificial Intelligence*, vol. 12, no. 1, pp. 1–12, Jan. 2000, https://doi.org/10.1080/095281300146272.

[18] T. Siddiqui, M. Latif, M. U. Farooq, M. A. Baig, and Y. S. Hassan, "Chronic Obstructive Pulmonary Disease Diagnosis with Bagging Ensemble Learning and ANN Classifiers," *Engineering, Technology & Applied Science Research*, vol. 14, no. 3, pp. 14741–14746, June 2024, https://doi.org/10.48084/etasr.7106.