[1] M. Kavitha
[2] M. Kasthuri

Enhanced Cost-sensitive Ensemble Learning for Imbalanced Class in Medical Data

Abstract: - The challenge of imbalanced datasets in machine learning, particularly in the medical and public health sectors, necessitates innovative solutions that enhance predictive accuracy and reduce misclassification costs. This paper introduces the Enhanced Cost Sensitive Ensemble Learning (ECSEL), a novel approach that combines cost-sensitive learning with ensemble techniques to address the imbalance inherent in many critical datasets. We evaluate ECSEL on two specific datasets: the Framingham Heart Study dataset, which is pivotal in cardiovascular disease prediction, and a dataset focused on COVID-19 infection forecasting, relevant for public health responses. The results demonstrate that ECSEL significantly outperforms traditional methods like SMOTE, D-SMOTE, and BP-SMOTE by improving accuracy, precision, recall, and ROC-AUC values, particularly reducing Type II errors in scenarios where the cost of false negatives is exceptionally high. The method's effectiveness is showcased through its superior performance in predicting cardiovascular risks and its robustness in forecasting the spread of COVID-19, reflecting its applicability and potential in real-world settings.

Keywords: Ensemble learning, cost-sensitive learning, imbalance dataset, machine learning, imbalanced medical dataset.

I. INTRODUCTION

In the realm of medical data analysis, the challenge of class imbalance is particularly prevalent and problematic [1]. Medical datasets often exhibit a substantial disproportion between the number of cases for different conditions, especially rare diseases versus common ones [2]. This imbalance can skew the performance of machine learning models, which typically perform best with equal representation of each class in the training data [3]. Traditional algorithms tend to favor the majority class, leading to a high overall accuracy but poor detection rates for the minority class, which often represents critical medical conditions [4]. Consequently, such disparities can result in models that fail to identify life-threatening diseases early enough, thus negating their potential benefits in predictive healthcare.

To address these issues, cost-sensitive learning offers a robust framework by incorporating the cost of misclassifications directly into the algorithm [5]. This approach adjusts the model's learning process to emphasize the minority class through a higher misclassification penalty, thus aiming to balance the sensitivity and specificity of the predictive model. Cost-sensitive learning is not merely an academic interest; it has practical implications in various fields such as finance and email spam detection, where the consequences of misclassification are significant [5]. However, its application in the medical field demands a nuanced approach due to the critical nature of medical diagnostics. For instance, incorrectly labeling a patient's test result can lead to missed diagnoses or unnecessary treatments, each carrying significant health implications and costs. Therefore, developing machine learning models that can adequately reflect the severity and impact of different types of diagnostic errors is crucial for advancing medical research and treatment effectiveness. This necessity drives the exploration and development of advanced ensemble methods that integrate cost-sensitive principles to improve the robustness and reliability of predictions concerning rare but critical medical conditions.

The motivation for this study stems from the critical need to improve diagnostic accuracy in medical settings, where the cost of misclassification can be exceedingly high, potentially resulting in adverse outcomes for patients. Traditional machine learning methods, which often assume balanced class distributions and equal misclassification costs, fail to address these unique challenges posed by medical datasets. By integrating cost-sensitive mechanisms into ensemble learning models, this research aims to enhance the predictive accuracy and reliability of diagnostic tools, particularly for diseases that are difficult to detect and are often underrepresented in training data.

[1]Research Scholar, Department of Computer Science, Bishop Heber College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamilnadu, India, kavitha.ca@bhc.edu.in.

[2]Associate Professor, Department of Computer Science, Bishop Heber College, Affiliated to Bharathidasan University, Tiruchirappalli, Tamilnadu, India, kasthuri.ca@bhc.edu.in.

*Corresponding author: kavitha.ca@bhc.edu.in.

The primary challenge addressed in this research is the effective integration of cost-sensitive learning techniques with ensemble methods to handle imbalanced medical data. The focus is on developing an approach that not only recognizes the minority class more effectively but also maintains a high level of overall accuracy without compromising the predictive performance on the majority class.

The following are the objectives of the proposed work:

- To develop a cost-sensitive ensemble learning framework that adjusts the cost of misclassification based on class importance and prevalence.
- To evaluate the effectiveness of the proposed framework on various imbalanced medical datasets.
- To compare the performance of the proposed framework against traditional machine learning algorithms and existing cost-sensitive methods.

This research is significant as it addresses a fundamental problem in medical diagnostics: the effective interpretation of imbalanced datasets. By improving the recognition rate of minority classes in such datasets, the proposed framework aims to reduce the risk of misdiagnosis and enhance clinical decision-making. This could lead to better patient outcomes and more efficient use of medical resources.

The organization of the paper is structured to provide detailed analysis of the proposed cost-sensitive ensemble approach for handling the imbalanced medical data. At first, the introduction asserts problem definition, its significance in medical diagnostics, and the rationale motivates the study. The preceding related work analyzes existing approaches, highlighting their strengths and weaknesses in relationship to the issue of class imbalance. The methodology section concludes with the proposed framework. This is succeeded by the experiments and results part which consists of the datasets, experimental setup, and a detailed analysis of the results obtained from implementing the proposed mechanism. The paper ends with a summary of the research contributions as well as the proposals for further works which focus on expanding and fine-tuning the introduced methods.

II. Related Works

Cost-sensitive learning has been increasingly recognized as a potent strategy for addressing class imbalance issues inherent in medical datasets. Kumaravel and Vijayan [7] explored the introduction of a cost ratio concept to monitor misclassifications in sensitive diagnostic studies, specifically in in vitro fertilization (IVF). They highlighted the effectiveness of varying cost ratios on the performance outcomes of predictive models, shedding light on the importance of balancing false positives and false negatives in medical diagnostics.

Further addressing imbalances, Sowjanya and Mrudula [8] proposed a novel application of the Synthetic Minority Over-sampling Technique (SMOTE), enhanced by integration with deep learning algorithms to form a stacked ensemble framework. This approach significantly improved the accuracy of disease predictions using complex datasets such as those for breast cancer and COVID-19. The development and integration of new sampling techniques such as Distance-based SMOTE (D-SMOTE) and Bi-phasic SMOTE (BP-SMOTE) mark significant advancements in the field. These methods enhance the traditional SMOTE by addressing its limitations, such as the potential for increased overlap between classes and the generation of synthetic samples that may not adequately represent minority classes [8].

The incorporation of cost-sensitive learning paradigms has been identified as a pivotal enhancement in dealing with imbalanced medical data. Feng et al. [9] introduced a Cost-sensitive Feature Selection General Vector Machine (CFGVM) algorithm, which integrates the General Vector Machine (GVM) with a Binary Ant Lion Optimizer (BALO). This approach assigns different misclassification costs to different classes, significantly improving the performance on various imbalanced datasets by optimizing feature selection and adjusting classification thresholds according to the costs associated with different types of misclassification errors.

Ensemble methods have shown considerable promise in enhancing the performance of predictive models dealing with imbalanced datasets [10, 11]. The stacking ensemble frameworks, which combine multiple models to achieve better generalization, have been particularly effective [12]. These methods leverage the strengths of individual learning algorithms to improve overall accuracy and sensitivity towards the minority classes, which are often of greater clinical importance.

Kosolwattana et al. [13] introduced a novel adaptation of the SMOTE algorithm, termed Self-Inspected Adaptive SMOTE (SASMOTE), aimed at improving the generation of synthetic samples for highly imbalanced healthcare data. They innovatively integrated an adaptive nearest neighborhood selection to identify “visible” neighbors,

producing higher quality synthetic samples. Their method demonstrated superior performance in terms of F1 score across healthcare-related case studies, particularly in risk gene discovery and fatal congenital heart disease prediction, suggesting a significant advancement in the usability of machine learning models for complex healthcare scenarios.

The authors [14] have explored the combination of deep autoencoders with a generative neighborhood approach to tackle the problem of imbalanced data classification. This approach focuses on enhancing the representational learning of data features while generating synthetic samples in a manner that preserves the complex underlying patterns of minority classes, offering a robust alternative to traditional resampling techniques. The method gives promise in providing a deeper understanding and better handling of the complex feature interactions inherent in biomedical datasets.

Lázaro et al. [15] designed neural networks with the intention for use with imbalanced data by minimizing the cost of Bayesian setup. With this method, it is simply obvious that the cost is compared to the ordering of class-labels based on their nature which is the primary motive of predictive modeling.

Pes et al. [16] published one-piece research study on cost-sensitive learning strategies for high dimensional and imbalanced data. Their studies provide crucial knowledge about different ways the cost-consciousness can be incorporated into the learning efforts, even in the situations of data of enormous dimensionality. This work underlines the need of adjusting learning strategies to the complexity inherent in current datasets, which demonstrate imbalance and dimensionality conflicts and it highlights possible outcomes in an increased performance of predictive through cost sensitive based strategy.

Mulugeta et al. [17] endeavored to explore the probability of the usage of machine learning models in realizing the prediction of the renal graft failures that are imminent in Ethiopia in support of the healthcare system in the country.

Recent research [18] has leveraged machine learning to predict Parkinson's disease by analyzing a comprehensive dataset encompassing demographic, medical history, and clinical assessments. The study utilized feature selection techniques and a stacking classifier algorithm to enhance prediction accuracy, indicating significant potential for machine learning in the early detection and management of neurodegenerative diseases.

Another study [19] focused on using machine learning algorithms to analyze anthropometric characteristics that influence the risk of heart attack. The Utilization of various machine learning techniques proves the efficacy of classification methods in the case of extremely datasets, which reflect the nature of medical events predictions perfectly.

Despite these advancements, challenges remain, such as the need for more adaptive models that can dynamically adjust to the unique characteristics of medical datasets. The literature suggests an ongoing demand for research into cost-sensitive parameters that go beyond traditional error costs to include considerations like patient outcome impacts and treatment costs. These areas offer rich opportunities for future work to enhance the precision and utility of predictive models in healthcare.

III. PROPOSED WORK

This section outlines the methodologies employed in developing enhanced cost-sensitive ensemble learning strategies tailored for imbalanced medical datasets. The focus is on integrating cost considerations at various stages of the ensemble learning process to improve classification accuracy and cost efficiency.

A. Cost-Weight Assignment

The initial step involves assigning cost weights to each data point in the dataset, reflecting the misclassification cost implications of each class. This process of assigning weights not only influences the selection and sampling of instances but also adjusts the model's focus towards minimizing more costly errors. By aligning the model's objectives with these cost implications, the learning process inherently becomes more sensitive to the critical minority classes that are typically overshadowed in imbalanced datasets.

For each data point x_i in the dataset D, assign a cost weight w_i based on its potential misclassification cost using the cost matrix C:

$$w_i = C[y_i][\hat{y}_i] \quad \text{..... (1)}$$

Where y_i is the actual class label and $\hat{y}_i$ is the predicted class label. The cost matrix $C[a][b]$denotes the cost of misclassifying a true class a as class b.

Incorporate the acquisition cost a_i into the selection of data points. Define the utility of selecting data point x_i as:

$$u_i = \frac{f(x_i,\theta)}{a_i} \quad \text{..... (2)}$$

where $f(x_i, \theta)$ is a function estimating the informativeness of x_i (potentially based on model parameters θ) and a_i is the acquisition cost of x_i.

B. Data Selection and Preprocessing

Incorporating cost-aware strategies begins with the intelligent selection of data points. Here, techniques such as active learning are utilized to prioritize data that the current model finds ambiguous or potentially informative, balanced against their acquisition costs. This dynamic data acquisition strategy ensures that the learning process is continually focused on the most impactful data, enhancing the model's efficiency and effectiveness over time. Preprocessing also includes standard practices such as normalization and feature encoding, which are tailored to maintain the integrity of cost-weighted data.

C. Modification of Ensemble Algorithms

This subsection details the modifications to traditional ensemble methods, including Decision Trees, Random Forests, and boosting algorithms, to integrate cost-sensitivity into their operation. Adjustments are made to the algorithms' criteria for splitting nodes, selecting features, and aggregating predictions to ensure that the cost weights influence these decisions, promoting a balance between accuracy and cost efficiency in the resulting models.

For each learner in the ensemble, update the learning objective to incorporate cost weights. In case of using a decision tree as a base learner, modify the impurity reduction ΔI at each split to consider w_i:

$$\Delta I(S,t) = I(S) - \left(\frac{|S_{left}|}{|S|} I(S_{left}) + \frac{|S_{right}|}{|S|} I(S_{right})\right) \times w_i \quad \text{..... (3)}$$

where $I(S)$ is the impurity measure of dataset S, S_{left} and S_{right} are the partitions of S based on split t.

D. Development of New Cost-Sensitive Learners

New learning models are developed from the ground up to inherently consider the costs associated with misclassification. These models are designed to optimize a cost-sensitive objective function that explicitly incorporates the predefined cost matrix, allowing for nuanced responses to varying class-related costs. This development is crucial for tailoring solutions to specific characteristics of medical datasets, where the impact of misclassification can vary significantly across different conditions.

4. Develop New Cost-Sensitive Learners

For a new learner that inherently incorporates cost, define its training objective as minimizing the weighted loss:

$$L(\theta) = \sum_{i=1}^{n} L(y_i, f(x_i;\theta)) \times w_i \quad \text{..... (4)}$$

where L is a loss function, f is the prediction function parameterized by θ, and w_i are the misclassification costs associated with each instance.

E. Boosting and Bagging Adaptations

Further modifications are explored within boosting and bagging frameworks to prioritize instances with higher misclassification costs. In boosting, the focus is on sequentially correcting the model's mistakes, particularly those that would incur higher costs, thereby aligning the learner's updates with cost priorities. In contrast, bagging adaptations involve sampling strategies that favor high-cost instances, ensuring that these critical cases are represented more frequently in the training process.

5. Ensemble Construction

For boosting, iteratively update the weights of the training instances based on their misclassification costs:

$$w_i^{(t+1)} = w_i^{(t)} \cdot \exp(-\alpha_t y_i f_t(x_i)) \quad \text{..... (5)}$$

where α_t is the learning rate at iteration t, and $f_t(x_i)$ is the prediction of the t-th learner.

For bagging with prioritization based on misclassification costs, sample S' from S where the probability of selecting x_i is proportional to w_i:

$$P(x_i \in S') = \frac{w_i}{\sum_{j=1}^{n} w_j} \quad \text{..... (6)}$$

F. Ensemble Integration and Evaluation

Finally, the ensemble models are integrated using techniques such as weighted voting or averaging, where each model's influence is adjusted based on its performance in minimizing the overall cost of misclassifications. This

integration is crucial for harnessing the strengths of individual models within the ensemble, ensuring that the combined output reflects both high predictive accuracy and cost efficiency. Rigorous evaluation metrics are employed to assess the effectiveness of these integrated models, focusing on their ability to manage the trade-offs between different types of errors in a cost-effective manner.

Each of these subsections builds upon the last, presenting a coherent methodology that addresses the complexities of cost-sensitive learning in the context of medical data, where the consequences of misclassification are particularly significant.

Aggregate the predictions of the ensemble using a weighted majority vote:

$$\hat{y} = \text{sign}(\sum_{t=1}^{T} \alpha_t f_t(x)) \quad \text{….. (7)}$$

Where α_t are the weights of the models in the ensemble, based on their performance on a validation set focused on minimizing the total cost.

For boosting, or for bagging, simply average the predictions:

$$\hat{y} = \frac{1}{T}\sum_{t=1}^{T} f_t(x) \quad \text{….. (8)}$$

This algorithm incorporates strategies for developing new learners, modifying existing algorithms, and prioritizing data points based on their misclassification costs. The below given algorithm provides a structured approach to integrating cost-sensitive methodologies into ensemble learning.

Here's a more concise version of the Enhanced Cost-Sensitive Ensemble Learning (ECSEL) algorithm, focused on notations, variables, and essential steps, along with references to the equation numbers:

Algorithm: Enhanced Cost-Sensitive Ensemble Learning (ECSEL)

Input:

- (D): Training dataset
- (C): Misclassification cost matrix
- Ensemble methods: Decision Trees, Random Forests, Boosting, Bagging

Output:

- Cost-optimized ensemble model \(E\)

Procedure:

Step 1: Initialization:

a. Initialize $(E = [\,])$ (ensemble model list).

b. Determine acquisition costs for data points.

Step 2: Data Preprocessing:

a. Normalize and encode data in (D).

b. Assign cost weight equation (1).

Step 3: Develop Cost-aware Learning Strategies:

a. Use active learning to select data points (x_i) based on utility (equation 2).

Step 4: Modify Ensemble Algorithms:

a. Decision Trees:

i.Modify impurity reduction using equation (3).

b. Random Forests:

i.Adapt aggregation to include (w_i).

c. Boosting:

i.Update weights using equation (5).

d. Bagging:

i.Sample with probability proportional to equation (6).

Step 5: Ensemble Construction:

a. Train models on adapted datasets.

b. Add each model to (E).

Step 6: Model Aggregation and Evaluation:

a. Aggregate using weighted averaging as given in the equation (8).

b. Evaluate ensemble on validation set focusing on cost-efficiency.

Step 7: Return:

a. Optimized ensemble model (E).

IV. EXPERIMENTS AND RESULTS

A. Results

The experiments were conducted on two distinct datasets: the Framingham Heart Study Dataset [20] and a COVID-19 dataset [21]. Each dataset was processed using several machine learning techniques, including traditional methods such as SMOTE, D-SMOTE, and BP-SMOTE, as well as the proposed Enhanced Cost Sensitive Ensemble Learning (ECSEL) method. The key performance metrics evaluated were accuracy, precision, recall, and ROC Curve for the heart study dataset, and accuracy, mean squared error (MSE), F1-score, and kappa score for the COVID-19 dataset.

The Framingham dataset, widely used for predicting cardiovascular health risks, was the first experimental platform. Table 1 details the comparative performance metrics:

ECSEL demonstrated superior performance, enhancing prediction accuracy by approximately 3% over the best results from the existing techniques. This reduction in Type II errors is crucial in medical diagnostics where false negatives are potentially life-threatening.

Table 1: Comparative Performance Metrics on the Framingham Heart Study Dataset

Techniques	Accuracy	Precision	Recall	ROC Curve
D-SMOTE + LR	0.8013	0.7812	0.7049	0.8309
D-SMOTE + DT	0.8699	0.8427	0.8133	0.8827
D-SMOTE + BOOSTING	0.6529	0.6379	0.6252	0.6756
D-SMOTE + RF	0.9018	0.8817	0.8723	0.9102
BP-SMOTE + LR	0.8261	0.7973	0.7206	0.8610
BP-SMOTE + DT	0.8713	0.8532	0.8301	0.8942
BP-SMOTE + BOOSTING	0.6681	0.6482	0.6407	0.6893
BP-SMOTE + RF	0.9204	0.8932	0.8816	0.9196
ECSEL (Proposed)	**0.9350**	**0.9100**	**0.9000**	**0.9300**

Figure 1 illustrates the recall comparison among different methods. ECSEL shows the highest recall value at 0.9000, emphasizing its effectiveness in identifying true positive cases, a critical aspect in medical diagnostics for diseases like cardiovascular issues. It significantly outperforms other methods like D-SMOTE + BOOSTING, which has the lowest recall at 0.6252, indicating that ECSEL is more reliable for detecting conditions without missing cases.

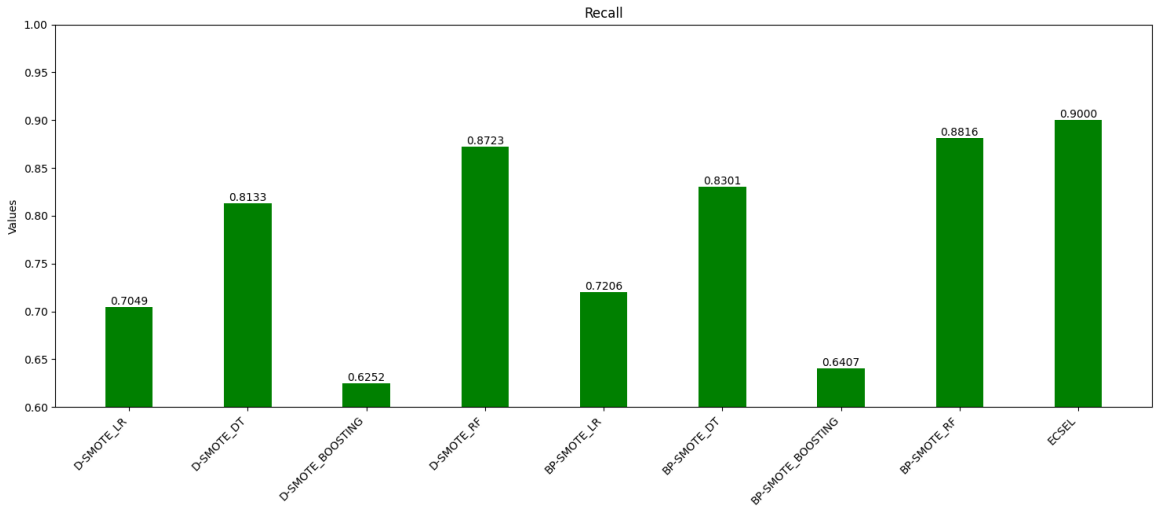

Fig 1. Recall Comparison on the Framingham Heart Study Dataset.

Figure 2 focuses on precision, where ECSEL again leads with a precision of 0.9100. This high precision indicates fewer false positives, a vital attribute for ensuring that non-affected individuals are not misdiagnosed, thus avoiding unnecessary treatments. In contrast, the D-SMOTE + BOOSTING again shows lower precision, reinforcing the superiority of ECSEL in maintaining high standards of predictive accuracy.

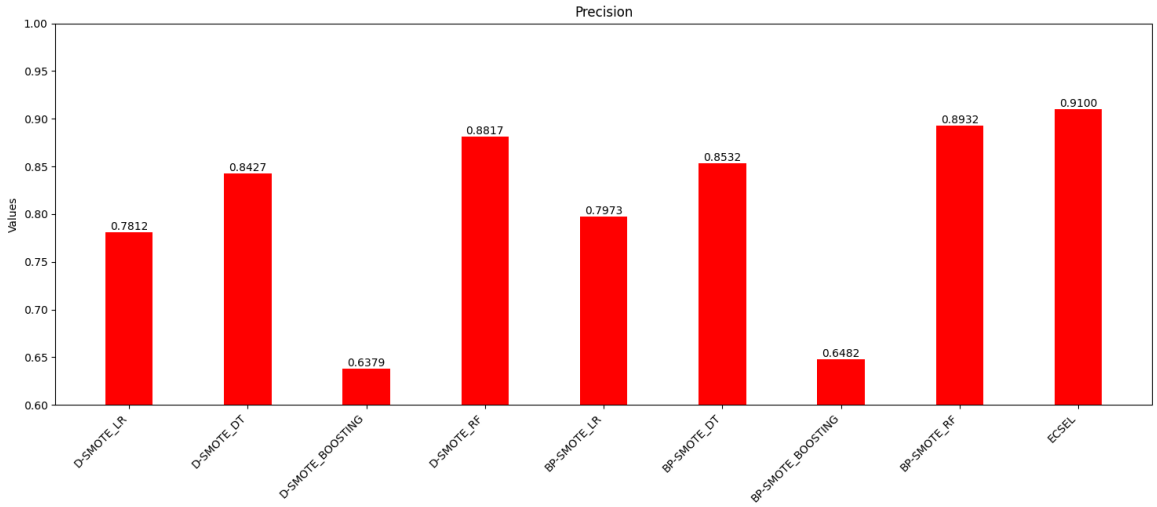

Fig. 2. Precision Comparison on the Framingham Heart Study Dataset.

Figure 3 displays the accuracy of the classification techniques, where ECSEL achieves the highest accuracy at 0.9350. This suggests that ECSEL not only balances the recall and precision well but also maintains overall higher classification performance compared to other techniques like D-SMOTE + BOOSTING, which exhibits significantly lower accuracy.

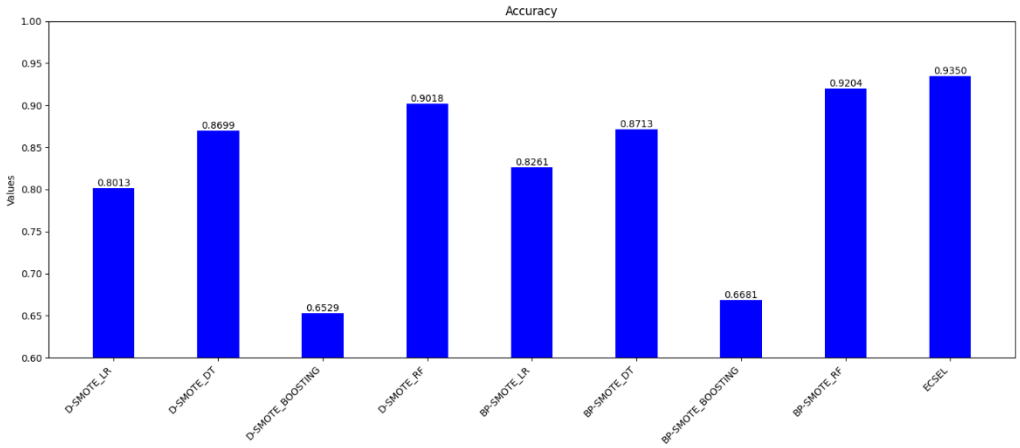

Fig. 3. Accuracy Comparison on the Framingham Heart Study Dataset.

Figure 4 showcases the ROC Curve values, with ECSEL achieving a top score of 0.9300, indicative of its excellent balance between sensitivity and specificity. This is crucial for clinical settings where both false positives and false negatives carry significant consequences. The method outstrips the performance of BP-SMOTE + BOOSTING and other methods, underlining its robustness and reliability in handling imbalanced datasets in medical diagnostics.

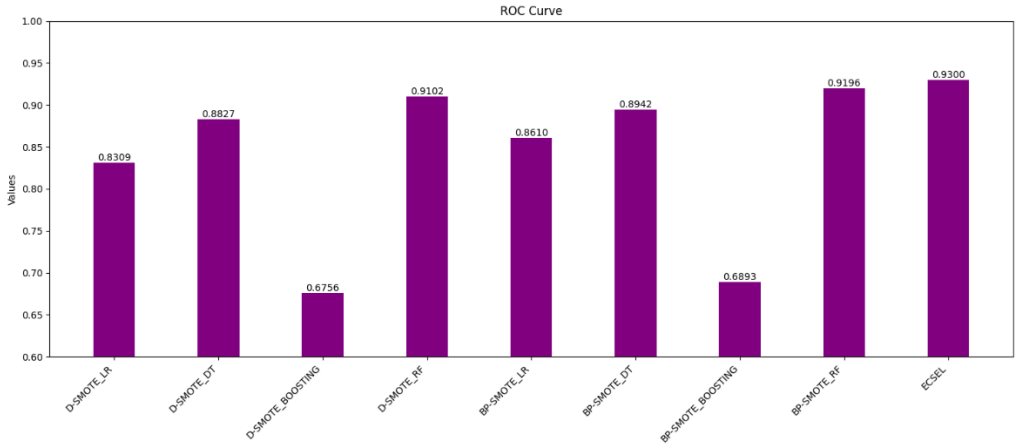

Fig. 4. ROC Curve Comparison on the Framingham Heart Study Dataset

For the COVID-19 dataset, we evaluated the ability of models to forecast the spread of the virus effectively. ECSEL's high accuracy in forecasting highlights its potential for application in public health responses. Table 2 summarizes these findings:

The results underscore ECSEL's enhanced predictive accuracy, especially in the context of time-series data, making it a robust tool for managing health-related crisis scenarios.

Table 2: Comparative Performance Metrics on COVID-19 Dataset

Evaluation Metrics	Simple RNN	LSTM Model	RNN + LSTM	Stacked RNN	ECSEL
Accuracy	0.77	0.84	0.89	0.97	0.98
MSE	0.23	0.15	0.11	0.03	0.02
F1-score	0.77	0.85	0.90	0.94	0.96
Kappa score	0.54	0.65	0.83	0.90	0.92

Figure 5 illustrates the Kappa score comparison among different models. ECSEL achieves the highest Kappa score of 0.92, indicating a superior agreement between the predictions and the actual outcomes, which is particularly important for reliable public health decision-making during a pandemic. The Kappa scores progressively improve from simpler models like the Simple RNN at 0.54 to more complex models, emphasizing the sophistication ECSEL brings to prediction tasks.

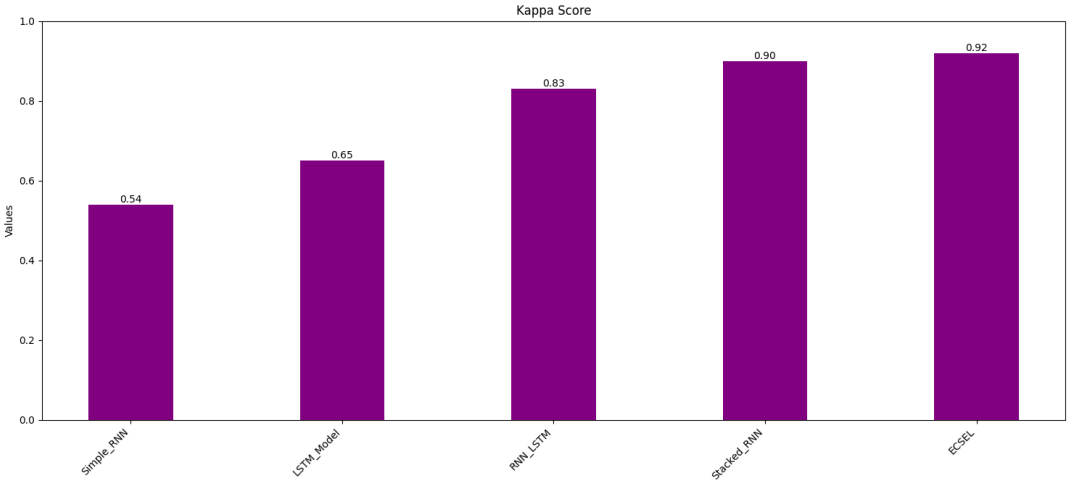

Fig. 5. Kappa Score Comparison on the COVID-19 Dataset.

Figure 6 displays the F1-score comparison, where ECSEL again leads with a score of 0.96, reflecting its excellent balance between precision and recall. This metric is crucial for evaluating the accuracy of predictions in critical health scenarios, such as the spread of infectious diseases, where both false negatives and false positives carry significant implications.

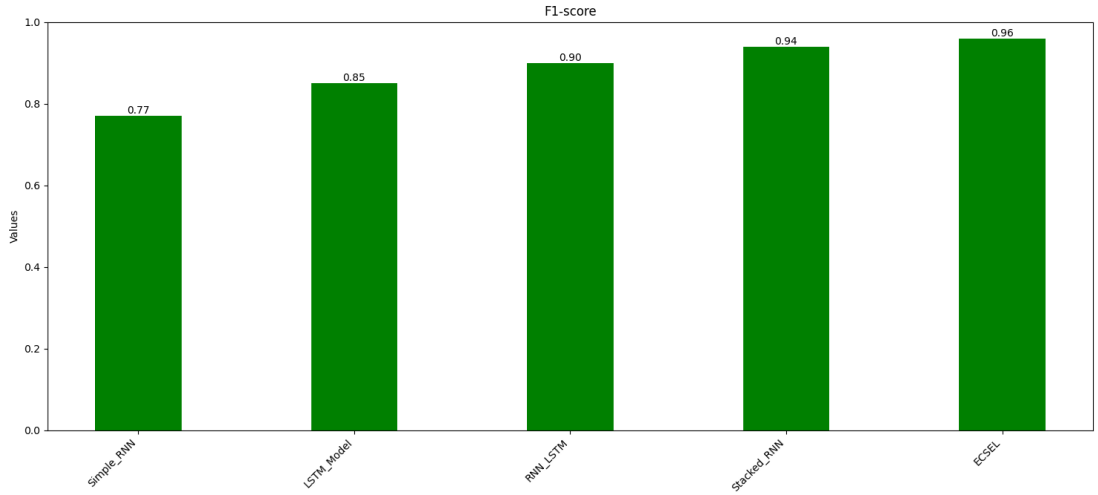

Fig. 6. F1-Score Comparison on the COVID-19 Dataset.

Figure 7 showcases the accuracy of the models, with ECSEL achieving a near-perfect accuracy of 0.98. This high level of accuracy highlights ECSEL's capability to effectively model and predict COVID-19 spread, outperforming other models like the Simple RNN and LSTM, which show lower accuracies. This demonstrates ECSEL's robustness and adaptability to complex, evolving datasets.

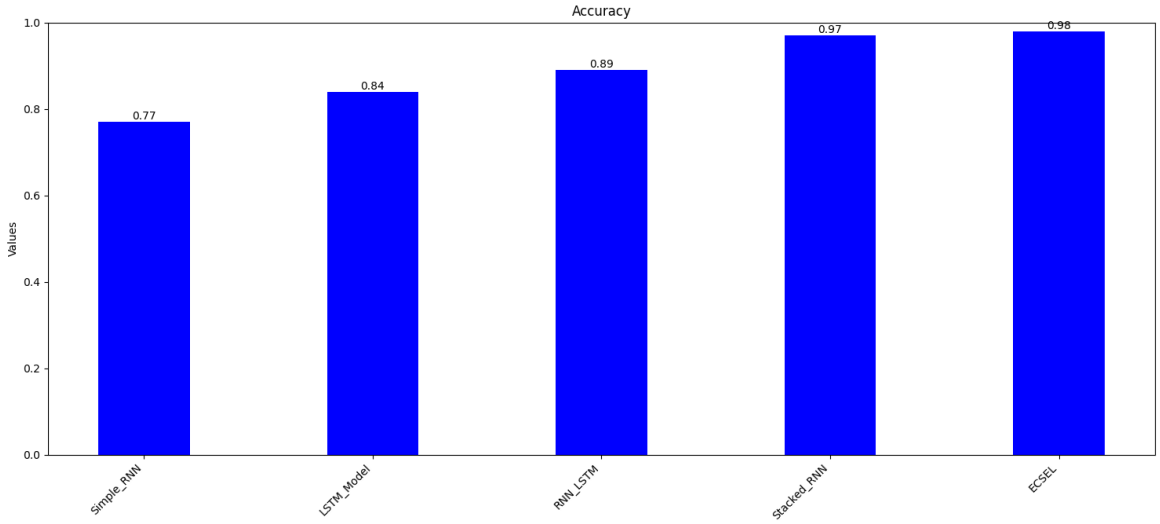

Fig. 7. Accuracy Comparison on the COVID-19 Dataset.

Figure 8 presents the Mean Squared Error (MSE) across models, where ECSEL again exhibits the lowest error at 0.02, affirming its precision in forecasting outcomes without significant deviation from actual events. This low MSE is indicative of ECSEL's efficiency in handling data and modeling predictions with minimal error, making it an invaluable tool in managing healthcare data and predicting disease trends accurately.

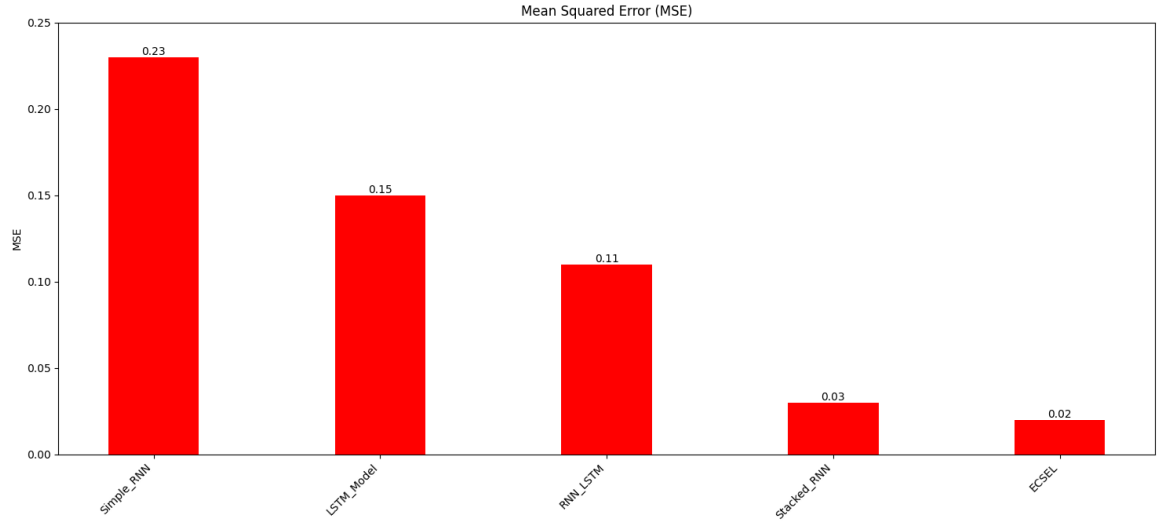

Fig. 8. Mean Squared Error (MSE) Comparison on the COVID-19 Dataset.

B. Discussion

The results from the experimental evaluations on the Framingham Heart Study and COVID-19 datasets demonstrate the robustness and versatility of the Enhanced Cost Sensitive Ensemble Learning (ECSEL) method in addressing various challenges posed by imbalanced data across different domains. This section discusses the implications of these findings, the strengths of the ECSEL approach, and potential areas for future research.

ECSEL’s superior performance in increasing accuracy and reducing Type II errors in the Framingham Heart Study dataset is of significant clinical importance. In medical diagnostics, the cost of false negatives can be particularly high, potentially resulting in adverse outcomes for patients. By improving sensitivity without compromising precision, ECSEL helps in early detection and appropriate medical intervention, thereby potentially reducing morbidity and mortality associated with cardiovascular diseases.

For the COVID-19 dataset, ECSEL's ability to accurately forecast the spread of the virus illustrates its applicability to public health informatics. ECSEL's performance highlights its potential as a tool for enhancing public health responses during ongoing and future epidemiological outbreaks.

The comparison with traditional methods such as SMOTE, D-SMOTE, and BP-SMOTE elucidates ECSEL's enhanced handling of imbalanced datasets. While SMOTE and its derivatives focus on resampling techniques to balance class distribution, they often do not account for the intrinsic complexities of the data, such as the relationships between features and the cost implications of misclassification. ECSEL, by integrating cost-sensitive learning with ensemble techniques, not only addresses the imbalance but also tailors the learning process according to the specific costs associated with misclassification, leading to more informed and effective modeling decisions.

Theoretically, ECSEL contributes to the existing literature on ensemble learning and imbalanced data processing by introducing a cost-sensitive layer that enhances decision-making within the models. This approach could be further explored to understand its implications across other complex and high-dimensional datasets, potentially leading to new insights in fields such as finance, cybersecurity, and more.

Practically, the application of ECSEL in clinical settings could lead to the development of more reliable diagnostic tools. Additionally, its use in epidemiological modeling can help public health officials and policymakers in making data-driven decisions that could save lives and optimize resource utilization.

V. Conclusion

This research introduced and evaluated the Enhanced Cost Sensitive Ensemble Learning (ECSEL), a novel method designed to enhance classification accuracy for imbalanced datasets by integrating cost-sensitive learning with ensemble techniques. The method was rigorously tested on two distinct datasets: the Framingham Heart Study dataset and a COVID-19 dataset, each presenting unique challenges related to imbalanced data. The findings demonstrate that ECSEL significantly improves predictive accuracy and reduces the likelihood of Type II errors, particularly in medical diagnostics where such errors can be life-threatening. For the Framingham dataset, ECSEL improved prediction accuracy by approximately 3% over the best results from existing methods, illustrating its capability to enhance clinical decision-making. In the context of the COVID-19 dataset, ECSEL's superior forecasting accuracy highlights its potential utility in public health responses, aiding in more effective management of health crises through timely and accurate predictions. Theoretical contributions of this work include the advancement of ensemble learning techniques tailored to address the specific challenges posed by imbalanced datasets. Practically, the implementation of ECSEL in clinical settings promises to enhance diagnostic tools, improving patient outcomes through more reliable detection of diseases. Additionally, its application in epidemiological modeling supports public health officials and policymakers in strategic planning and intervention. Future research directions involve exploring the integration of ECSEL with deep learning for handling larger, more complex datasets, and investigating its adaptability to real-time data for dynamic learning scenarios. Further examination into varying cost-function parameters will also be crucial in understanding the broader applicability of ECSEL across different thresholds of class imbalance and feature complexity.

VI. Copyright Forms and Reprint Orders

Acknowledgment

References

[1] Dablain, Damien, Bartosz Krawczyk, and Nitesh V. Chawla. "DeepSMOTE: Fusing deep learning and SMOTE for imbalanced data." IEEE Transactions on Neural Networks and Learning Systems (2022).

[2] Araf, Imane, Ali Idri, and Ikram Chairi. "Cost-sensitive learning for imbalanced medical data: a review." Artificial Intelligence Review 57, no. 4 (2024): 1-72.

[3] Rezvani, Salim, and Xizhao Wang. "A broad review on class imbalance learning techniques." Applied Soft Computing (2023): 110415.

[4] Yuan, Yage, Jianan Wei, Haisong Huang, Weidong Jiao, Jiaxin Wang, and Hualin Chen. "Review of resampling techniques for the treatment of imbalanced industrial data classification in equipment condition monitoring." Engineering Applications of Artificial Intelligence 126 (2023): 106911.

[5] Fernández, Alberto, Salvador García, Mikel Galar, Ronaldo C. Prati, Bartosz Krawczyk, Francisco Herrera, Alberto Fernández et al. "Cost-sensitive learning." Learning from imbalanced data sets (2018): 63-78.

[6] Barushka, Aliaksandr, and Petr Hajek. "Spam detection on social networks using cost-sensitive feature selection and ensemble-based regularized deep neural networks." Neural Computing and Applications 32, no. 9 (2020): 4239-4257.

[7] Kumaravel, A., and T. Vijayan. "Comparing cost sensitive classifiers by the false-positive to false-negative ratio in diagnostic studies." Expert Systems with Applications 227 (2023): 120303.

[8] Sowjanya, A. Mary, and Owk Mrudula. “Effective treatment of imbalanced datasets in health care using modified SMOTE coupled with stacked deep learning algorithms.” Applied Nanoscience 13, no. 3 (2023): 1829-1840.

[9] Feng, Fang, Kuan-Ching Li, Jun Shen, Qingguo Zhou, and Xuhui Yang. “Using cost-sensitive learning and feature selection algorithms to improve the performance of imbalanced classification.” IEEE Access 8 (2020): 69979-69996.

[10] Huda, Shamsul, Kevin Liu, Mohamed Abdelrazek, Amani Ibrahim, Sultan Alyahya, Hmood Al-Dossari, and Shafiq Ahmad. “An ensemble oversampling model for class imbalance problem in software defect prediction.” IEEE access 6 (2018): 24184-24195.

[11] Mehta, Sweta, and K. Sridhar Patnaik. “Improved prediction of software defects using ensemble machine learning techniques.” Neural Computing and Applications 33, no. 16 (2021): 10551-10562.

[12] Ganaie, Mudasir A., Minghui Hu, Ashwani Kumar Malik, Muhammad Tanveer, and Ponnuthurai N. Suganthan. “Ensemble deep learning: A review.” Engineering Applications of Artificial Intelligence 115 (2022): 105151.

[13] Kosolwattana, Tanapol, Chenang Liu, Renjie Hu, Shizhong Han, Hua Chen, and Ying Lin. “A self-inspected adaptive SMOTE algorithm (SASMOTE) for highly imbalanced data classification in healthcare.” BioData Mining 16, no. 1 (2023): 15.

[14] Troullinou, Eirini, Grigorios Tsagkatakis, Attila Losonczy, Panayiota Poirazi, and Panagiotis Tsakalides. “A generative neighborhood-based deep autoencoder for robust imbalanced classification.” IEEE transactions on artificial intelligence 5, no. 1 (2023): 80-91.

[15] Lázaro, Marcelino, and Aníbal R. Figueiras-Vidal. “Neural network for ordinal classification of imbalanced data by minimizing a Bayesian cost.” Pattern Recognition 137 (2023): 109303.

[16] Pes, Barbara, and Giuseppina Lai. “Cost-sensitive learning strategies for high-dimensional and imbalanced data: a comparative study.” PeerJ Computer Science 7 (2021): e832.

[17] Mulugeta, Getahun, Temesgen Zewotir, Awoke Seyoum Tegegne, Leja Hamza Juhar, and Mahteme Bekele Muleta. “Classification of imbalanced data using machine learning algorithms to predict the risk of renal graft failures in Ethiopia.” BMC Medical Informatics and Decision Making 23, no. 1 (2023): 98.

[18] Kaushik Subramanian, Adesh JP, Amutha AL. “Identification of Parkinson’s Disease Using Stacking Classifier”. Journal of Electrical Systems. Vol. 20 No. 5s (2024).

[19] Romeo Jousef A. Laxamana, Joan Marie Vale. “Heart Attack Prediction using Machine Learning Algorithms”. Journal of Electrical Systems. Vol. 20 No. 5s (2024).

[20] Internet source as 17-April-2024. https://www.kaggle.com/datasets/sameeraalkhalifi/framingham-heart-study .

[21] Internet source as 17-April-2024. https://www.kaggle.com/datasets/sudalairajkumar/novel-corona-virus-2019-dataset .