IJCRT.ORG ISSN : 2320-2882

INTERNATIONAL JOURNAL OF CREATIVE RESEARCH THOUGHTS (IJCRT)

An International Open Access, Peer-reviewed, Refereed Journal

“Performance Evaluation Of Tree-Based And Deep Neural Models For STD Classification: A Comparative Approach”

[1]Aaysha Sheikh, [2]Shruti Kondekar, [3]Iqra Siddiqui, [4]Saniya Sheikh

Mrunali Vaidya, Professor
Computer Science Engineering
Ballarpur Institute of Technology, Ballarpur, India

Abstract: Early detection of sexually transmitted diseases (STDs) and infections (STIs) is crucial for timely treatment and reducing public health risks. In the first phase of this study, a Random Forest model was applied on an imbalanced STD dataset and achieved an accuracy of 96%, demonstrating the potential of tree-based learning in medical diagnostics. In the second phase, we addressed data imbalance using the Synthetic Minority Oversampling Technique (SMOTE) and extended the analysis by implementing deep learning models such as Multilayer Perceptron (MLP) and 1D Convolutional Neural Network (1D CNN), along with Gradient Boosting. A comprehensive performance evaluation was conducted using accuracy, precision, recall, F1-score, and ROC-AUC. Among all, Random Forest achieved the highest accuracy of 94% on the resampled data, highlighting the promise of deep learning in enhancing diagnostic accuracy. This final phase reinforces the effectiveness of combining data balancing techniques with advanced learning models for robust STD prediction.

Index Terms - Tree Learning, Deep learning, Sexually Transmitted Disease (STD) Prediction, Random Forest, MLP, Gradient Boosting, 1D CNN, SMOTE, Deep Learning, Machine Learning, Evaluation Metrics.

I. INTRODUCTION

Sexually transmitted diseases (STDs) and infections (STIs) pose a persistent global health burden, especially in regions with limited access to timely diagnostics and treatment. Early detection plays a crucial role in reducing the risk of complications, preventing transmission, and improving public health outcomes. While conventional testing methods remain effective, they often require clinical infrastructure and lab-based analysis, which can delay diagnosis and treatment.

To overcome these limitations, artificial intelligence (AI) and machine learning (ML) approaches are increasingly being explored for medical diagnostics. In Phase I of this study, we implemented a tree-based model—Random Forest—on a publicly available STD dataset. This model achieved high predictive performance, with an accuracy of 96%, and showed balanced classification capability for both positive and negative cases. Despite the dataset being relatively balanced (with only a small class difference), we recognized the potential for even slight imbalances to affect certain classifiers under specific conditions.

As we progressed into Phase II, we focused on further strengthening model robustness and exploring more advanced learning architectures. To ensure fair representation across both classes, we applied Synthetic Minority Oversampling Technique (SMOTE) to produce a more uniformly distributed dataset. While the original class imbalance was minor, SMOTE allowed us to validate how data balancing might influence performance across different algorithmic families.

Following this, we implemented and compared models from both tree-based (Gradient Boosting) and deep learning (Multilayer Perceptron and 1D Convolutional Neural Network) paradigms. Each model was trained on the resampled dataset and evaluated using comprehensive metrics including accuracy, confusion matrix, precision, recall, F1-score, and ROC-AUC.

Among all, the MLP model delivered the best overall performance on the resampled data, illustrating the potential of deep learning methods in healthcare diagnostics—even with tabular data. The CNN model, while slightly less accurate, still demonstrated the ability to learn meaningful representations. Gradient Boosting provided a competitive benchmark among the tree-based techniques.

This paper presents the findings of Phase II and offers a final comparative perspective between tree-based and deep learning models. Together with the insights from Phase I, our work contributes toward the development of a reliable, intelligent system for early STD/STI detection using AI technologies.

II. OBJECTIVE

Building upon the foundation established in Phase I, the objective of Phase II is to extend the comparative analysis by incorporating advanced machine learning and deep learning techniques on a more robust dataset. This phase specifically focuses on evaluating how data balancing (via SMOTE) and complex model architectures influence prediction performance for STD detection. The key objectives of this phase are:

1. **Comparison Enhancement of Model Diversity and Complexity:** Assess Expand the scope of models by implementing advanced algorithms such as Gradient Boosting (tree-based) and deep learning architectures like Multilayer Perceptron (MLP) and 1D Convolutional Neural Network (1D-CNN). These models are evaluated for their ability to handle tabular health data and identify non-linear patterns in the diagnosis of STDs.
2. **Integration of Data Balancing Techniques:** Apply SMOTE (Synthetic Minority Over-sampling Technique) to the dataset—even with only a slight class imbalance—to improve model generalizability and verify how resampling impacts model performance across both algorithm families.
3. **Comparative Evaluation Using Standard Metrics:** Systematically compare all models (Random Forest, Gradient Boosting, MLP, and 1D-CNN) on resampled data using metrics such as accuracy, precision, recall, F1-score, ROC-AUC, and confusion matrix, providing a holistic view of each model’s diagnostic effectiveness.
4. **Analysis of Trade-offs and Practical Implications:** Evaluate trade-offs between traditional and deep learning methods in terms of computational cost, interpretability, training complexity, and suitability for real-world healthcare deployment.

III. CONTRIBUTION

1. **Integration of Deep Learning Models**: Expanded the study beyond traditional machine learning by implementing MLP and 1D-CNN, demonstrating the viability of deep learning techniques for tabular health data.
2. **Data Resampling and Fair Evaluation**: Applied SMOTE even for a mildly imbalanced dataset to ensure robustness and investigate its impact on model performance.
3. **Expanded Model Comparison Framework** Added Gradient Boosting to the tree-based category, and conducted a four-way comparative analysis between Random Forest, Gradient Boosting, MLP, and 1D-CNN.
4. **Model Evaluation Under Realistic Conditions**: Compared model behaviour before and after data balancing, providing insight into how minor class imbalance affects different model types.
5. **Visual and Metric-Based Validation**: Used ROC curves, AUC scores, confusion matrices, and classification reports to thoroughly validate each model’s reliability and practical applicability in healthcare diagnostics.

IV. Literature Review

The reviewed studies present a range of methodologies, issues, and limitations in the domain of machine learning for sexually transmitted infection (STI) prediction. Methodologies employed include data collection from clinical records, self-reported surveys, and electronic health records [9][1], with preprocessing techniques like one-hot encoding, feature selection based on expert opinions, and stratified sampling to ensure data balance [9]. Machine learning models such as logistic regression, ensemble methods, MySTIRisk[1][2], and CatBoost[5] were evaluated using metrics like AUC, sensitivity, and F1-score, with some leveraging tools like Youden's Index to optimize diagnostic thresholds. However, common issues include biases arising from self-reported data, recall errors, and social desirability factors, as well as inconsistencies in data collection and definitions of risk behaviours across studies. Limitations are evident in the generalizability of findings due to the use of specific high-risk populations or single-site datasets, the exclusion of critical socio-behavioural and structural determinants, and the lack of external validation of predictive models. Furthermore, publication bias, restricted access to resources, and the absence of standardized methodologies hinder comprehensive comparisons and broader applicability. Addressing these challenges requires the inclusion of diverse datasets, integration of clinical images, and cross-site validation to improve the robustness and reliability of STI prediction models[2].

Year	Methodologies	Accuracy
2022	Boosted GLM, ensemble Elastic-Net Regression, Gradient Boosting Machines (GBM), Random Forest (RF)	67-76%
2023	Random Forest, Naïve Bayes classifier, and Decision Tree, SVM with three different kernel and Logistic Regression	65-70%
2024	logistic regression	68-69%
2024	-	56%
2024	CatBoost, Gradient Boost, Light GBM, SVM, Adaboost Classifier, Extinct Gradient Boosting, Logistic Regression, Random Forest	83%

V. METHODOLOGY

Phase I: Tree-Based Learning (Random Forest)

1. **Dataset Preparation**: Collected a publicly available dataset related to sexually transmitted diseases (STDs). Performed data cleaning, preprocessing, and label encoding to handle categorical variables. Divided the data into training and testing subsets using an 80:20 split.
2. **Model Implementation – Random Forest**: Trained a Random Forest classifier as the baseline model for Phase I. Used default hyperparameters with stratified sampling to preserve class distribution. The model was trained on the original dataset (without any resampling techniques).
3. **Model Evaluation**: Evaluated using metrics such as accuracy, precision, recall, F1-score, and confusion matrix. Achieved high accuracy of 96%, demonstrating strong performance on imbalanced data. Feature importance analysis was also conducted to understand key predictors.

Phase II: Advanced Modeling and Deep Learning

I. Application of SMOTE (Synthetic Minority Oversampling Technique)

In Phase II of our research, we applied SMOTE to address the minor imbalance present in our dataset. Although the difference between positive and negative cases was relatively small (approximately 6 data points), SMOTE was employed to ensure fair training conditions for all machine learning models.

Why SMOTE?

i) Traditional models may develop a slight bias toward the majority class, even with small imbalances.

ii) SMOTE generates synthetic data points for the minority class rather than duplicating existing data, leading to better generalization.

iii) It helps ensure that performance metrics like recall and F1-score remain high for both classes, not just the majority.

Steps Involved:

1. **Detection of Class Imbalance:**

- Initial dataset had a slight skew in class distribution.

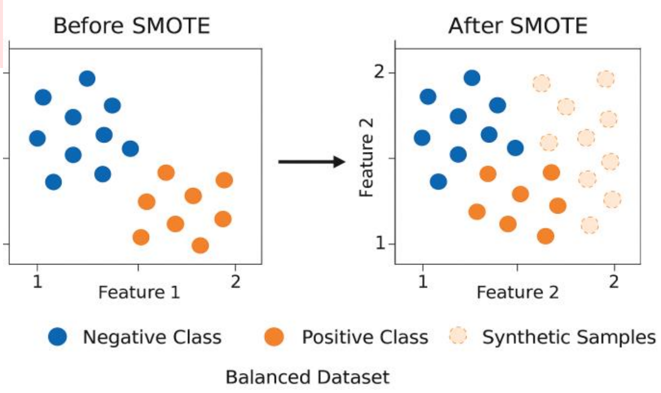
2. **Application of SMOTE:**

- Used SMOTE to synthetically generate new samples for the minority class.

- Resulted in a balanced dataset with equal representation of both classes.

3. **Model Training on Resampled Data:**

- Models such as Random Forest, Gradient Boosting, MLP, and 1D-CNN were trained using this balanced dataset.



II. Implementation of MLP (Multi-layer Perceptron)

What is MLP?

a. MLP is a type of feedforward neural network that consists of an input layer, one or more hidden layers, and an output layer.
b. It is used for supervised learning tasks and is capable of capturing complex non-linear patterns in data.

MLP Architecture Used in Our Study

1. Input Layer:

Takes in the feature values (after encoding and scaling).

Number of neurons = number of input features.

2. Hidden Layer(s):

We used two hidden layers first with 75 neurons and second with 20 neurons.

We used the ReLU activation function for both the hidden layers.

Why ReLU? It avoids the vanishing gradient problem and is computationally efficient.

3. Output Layer:

A single neuron with sigmoid activation to output probability for binary classification.

Why Sigmoid? It converts output into a probability (0 to 1), ideal for binary outcomes (Positive/Negative STD).

4. Training:

Loss Function: Binary Crossentropy – suitable for binary classification.

Optimizer: Adam – efficient for deep learning tasks.

Epochs: 2000 (early stopping was also applied later to avoid overfitting).

Scaler: StandardScaler was used to normalize data.

Early stopping was enabled to prevent overfitting, and the maximum number of iterations was set to 2000.

III. Implementation of Gradient Boosting

What is Gradient Boosting?

a. Gradient Boosting is an ensemble machine learning technique that builds models sequentially.
b. Each new model attempts to correct the errors made by the previous ones.
c. It combines the predictions of multiple weak learners (typically decision trees) to form a strong learner.

How It Works

1. Starts with a simple model (like a shallow tree).
2. Calculates the errors (residuals) of the model.
3. Builds a new model to predict these residuals.
4. Combines the new model with the previous one.
5. Repeats the process for several iterations (boosting).

Model Parameters (Default in our case):

Parameter	Description	Typical Value / Our Setting
N estimator	Number of boosting stages	Default (100)
Learning rate	Shrinks the contribution of each tree	Default (0.1)
Max Depth	Maximum depth of individual trees	Default (3)
Subsample	Fraction of samples used per tree	Default (1.0)
Loss	Loss function used	Default ('deviance')

Why Use Gradient Boosting?

Handles complex patterns better than a single decision tree.
More resistant to overfitting than plain boosting.
Offers better performance on tabular data.

IV.1D CNN model

A 1D Convolutional Neural Network is a deep learning model that is especially effective for analyzing sequential or structured data (like time series or tabular datasets).

It learns patterns by applying filters over one-dimensional feature sequences.

1D CNN Architecture Used in Our Model

Layer	Configuration	Purpose
Input Layer	Reshaped data: (samples, features, 1)	Prepares data for 1D conv operations
Convolution 1D layer	32 filters, kernel size = 3, ReLU	Extracts local feature patterns
MaxPooling 1D	Pool size = 2	Reduces dimensionality, retains key features
Dropout	30%	Prevents overfitting
Flatten		Converts 3D output to 1D
Dense layer	34 units, ReLU	Fully connected layer
Dropout	30%	Extra regularization
Output	1 unit, Sigmoid	Binary classification (STD/No STD)

Hyperparameters

a. Optimizer: Adam
b. Loss Function: Binary Crossentropy
c. Epochs: 220
d. Batch Size: 100

Why 1D CNN?

Captures feature interactions better than traditional models.
Automatically extracts features without manual engineering.
Good for structured/tabular datasets when reshaped appropriately.

V.Model Training Process – Phase II

This section outlines the common training workflow we followed for all models (MLP, Gradient Boosting, 1D CNN):

Step-by-Step Training Pipeline

1. Load Dataset:

- Resampled dataset using SMOTE to balance the classes.

2. Preprocessing

- Label encoding for categorical variables.
- Feature scaling using StandardScaler to normalize inputs.

3. Train-Test Split

- Dataset split into 80% training and 20% testing using `train_test_split`.

4. Model Building

- MLP: Fully connected neural network with hidden layers.
- Gradient Boosting: Ensemble of decision trees with sequential learning.
- 1D CNN: Deep learning model with convolution and pooling layers.

5. Model Compilation (for Deep Learning models)

- Optimizer: Adam
- Loss Function: Binary Crossentropy
- Metrics: Accuracy

6. Model Training

- MLP & CNN trained for multiple epochs with batch size, validation split, and in some cases, early stopping.
- Gradient Boosting trained using `GradientBoostingClassifier`.

7. Model Evaluation

- Metrics used: Accuracy, Precision, Recall, F1-Score, Confusion Matrix, AUC-ROC.

VI. Model Evaluation Metrics (Accuracy, Precision, Recall, F1 Score, Confusion Matrix, ROC-AUC)

After training, each model was rigorously evaluated to ensure reliability, robustness, and generalizability.

Evaluation Techniques

To assess the models' performance, we employed several evaluation metrics:

1. **Accuracy**: Measures the proportion of correctly predicted instances out of the total instances. Useful for balanced datasets.

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

2. **Precision**: Indicates the proportion of true positive predictions out of all positive predictions, highlighting the model's reliability in identifying true positives.

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

3. **Recall (Sensitivity)**: Measures the model's ability to identify all relevant instances, especially critical in healthcare applications.

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

4. **F1-Score**: The harmonic means of precision and recall, providing a balanced measure of model performance.

$$\text{F1-Score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

5. **Confusion Matrix**: A tabular representation of true positives (TP), true negatives (TN), false positives (FP), and false negatives (FN), offering a detailed breakdown of classification performance.

6. **ROC-AUC (Receiver Operating Characteristic - Area Under Curve)**: Evaluates the model's ability to distinguish between classes across different thresholds, with higher AUC indicating better discrimination.

7. **Cross-Validation (K-Fold)**: Each model was evaluated using K-Fold Cross Validation to assess consistency across multiple data splits. Reduces the risk of overfitting and ensures the model generalizes well.

Key Results

We repeated the training process few times and observe the accuracy is on range given below:

Model	Accuracy
Random Forest	**0.87 – 0.94**
Gradient Boosting	**0.80 – 0.84**
MLP	**0.70 – 0.73**
1D CNN	**0.70 – 0.76**

VII. Result and Conclusion

This study extended by introducing advanced machine learning and deep learning models, evaluated on a resampled STD/STI dataset using SMOTE to ensure class balance.
All models were also validated using cross-validation, which confirmed the consistency and generalizability of the performance metrics.

Conclusion:

1. **SMOTE balancing** helped mitigate minor class imbalance and ensured fairer model training.
2. **Random Forest** emerged as the top-performing model, achieving an impressive accuracy of 92% after resampling.
3. **Gradient Boosting** and **1D CNN** showed competitive performance but slightly underperformed compared to Random Forest.
4. **MLP**, despite being a deep learning model, did not perform as expected — likely due to the dataset size and complexity being insufficient for convolutional layers to learn effectively.
5. **Tree-based methods (Random Forest and Gradient Boosting)** remain strong, especially with smaller or structured datasets, but 1D-CNN demonstrated the potential of deep learning when tuned well.

VIII. Future Scope

- Experiment with more advanced architectures (e.g., LSTM, hybrid models).
- Apply feature engineering to extract domain-relevant insights.
- Extend analysis to larger, more diverse datasets to validate deep learning advantages.

IX. Reference

1. Melbourne Sexual Health Centre. MySTIRisk 2023. https://mystirisk.mshc.org.au/.8. Xu XL, Ge ZY, Chow EPF, Yu Z, Lee D, Wu JR, et al. A machine-learning-based risk-prediction tool for HIV and sexually transmitted infections acquisition over the next 12 months. J Clin Med 2022;11(7).

2. Latt PM, Soe NN, Xu X, Ong JJ, Chow EPF, Fairley CK, Zhang L. Identifying individuals at high risk for HIV and sexually transmitted infections with an artificial intelligence–based risk assessment tool. J Clin Epidemiol. 2024.

3. Chen Y, Yu W, Cai L, Liu B, Guo F. Enhancing HIV/STI decision-making: challenges and opportunities in leveraging predictive models for individuals, healthcare providers, and policymakers. 2024 22:886 https://doi.org/10.1186/s12967-024-05684-9

4. Kassaw AA, Yilma TM, Sebastian Y, Birhanu AY, Melaku MS, Jemal SS. Spatial distribution and machine learning prediction of sexually transmitted infections and associated factors among sexually active men and women in Ethiopia, evidence from EDHS 2016. BMC Infect Dis. 2023; 23:49. doi:10.1186/s12879-023-07987-6.

5. Soe NN, Latt PM, Yu Z, Lee D, Kim CM, Tran D, Ong JJ, Ge Z, Fairley CK, Zhang L. Clinical features-based machine learning models to separate sexually transmitted infections from other skin diagnoses. J Infect. 2024

6. Jiang Z, Xiu C, Yang J, Zhang X, Liu M, Chen X, et al. (2018) HIV test uptake and related factors amongst heterosexual drug users in Shandong province, China. PLoS ONE 13(10):e0204489. https://doi.org/10.1371/journal.pone.0204489

7. Wei C, Herrick A, Raymond HF, Anglemyer A, Gerbase A, Noar SM. Social marketing interventions to increase HIV/STI testing uptake among men who have sex with men and male-to-female transgender women. Cochrane Database of Systematic Reviews 2011, Issue 9. Art. No.: CD009337. DOI: 10.1002/14651858.CD009337

8. Ooi CY, Ng CJ, Sales AE, Lim HM Implementation Strategies for Web-Based Apps for Screening: Scoping Review J Med Internet Res 2020;22(7):e15591 URL: http://www.jmir.org/2020/7/e15591/ doi: 10.2196/15591

9. Farran B, AlWotayan R, Alkandari H, Al-Abdulrazzaq D, Channanath A and Thanaraj TA (2019) Use of Non-invasive Parameters and Machine-Learning Algorithms for Predicting Future Risk of Type 2 Diabetes: A Retrospective Cohort Study of Health Data From Kuwait. Front. Endocrinol. 10:624. doi: 10.3389/fendo.2019.00624

10. Park JH, Cho HE, Kim JH, Wall MM, Stern Y, Lim H, Yoo S, Kim HS, Cha J. Machine learning prediction of incidence of Alzheimer’s disease using large-scale administrative health data. npj Digit Med. 2020;3:46. doi:10.1038/s41746-020-0256-0.

11. Haakenstad A, Moses MW, Tao T, Tsakalos G, Zlavog B, Kates J, Wexler A, Murray CJL, Dieleman JL. Potential for additional government spending on HIV/AIDS in 137 low-income and middle-income countries: an economic modelling study. Lancet HIV. 2019;6:e382–e395. doi:10.1016/S2352-3018(19)30038-4.

12. Zhao PZ, Wang YJ, Cheng HH, Zhang Y, Tang WM, Yang F, Zhang W, Zhou JY, Wang C. Uptake and correlates of chlamydia and gonorrhea testing among female sex workers in Southern China: a cross-sectional study. BMC Public Health. 2021;21:1477. https://doi.org/10.1186/s12889-021-11526-w.

13. Falasinnu T, Gustafson P, Hottes TS, Gilbert M, Ogilvie G, Shoveller J. A critical appraisal of risk models for predicting sexually transmitted infections. Sex Transm Dis. 2014 May;41(5):321-30. doi: 10.1097/OLQ.0000000000000120. PMID: 24722388.

14. Scott H, Vittinghoff E, Irvin R, et al. Development and validation of the personalized sexual health promotion (SexPro) HIV risk prediction model for men who have sex with men in the United States. AIDS Behav 2020; 24:274–83.

15. Hoenigl M, Weibel N, Mehta SR, et al. Development and validation of the San Diego early test score to predict acute and early HIV infection risk in men who have sex with men. Clin Infect Dis 2015; 61:468–75.

16. Nieuwenburg SA, Hoornenborg E, Davidovich U, de Vries HJC, Schim van der Loeff M. Developing a symptoms-based risk score for infectious syphilis among men who have sex with men. Sex Transm Infect. 2023 Aug;99(5):324-329. doi: 10.1136/sextrans-2022-055550. Epub 2022 Nov 18. PMID: 36400527; PMCID: PMC10359546.

17. Ying GS, Maguire MG, Glynn RJ, Rosner B. Tutorial on Biostatistics: Receiver-Operating Characteristic (ROC) Analysis for Correlated Eye Data. Ophthalmic Epidemiol. 2022 Apr;29(2):117-127. doi: 10.1080/09286586.2021.1921226. Epub 2021 May 12. PMID: 33977829; PMCID: PMC8586066.

18. Ruiz MS, O'Rourke A, Allen ST, Holtgrave DR, Metzger D, Benitez J, Brady KA, Chaulk CP, Wen LS. Using interrupted Time Series Analysis to measure the impact of Legalized Syringe Exchange on HIV diagnoses in Baltimore and Philadelphia. J Acquir Immune Defic Syndr. 2019;82(Suppl 2):S148–54. https://doi.org/10.1097/qai.0000000000002176.

19. Ong JJ, Peng MH, Wong WW, Lo Y-R, Kidd MR, Roland M, Zhu SZ, Jiang SF. Opportunities and barriers for providing HIV testing through community health centers in mainland China: a nationwide cross-sectional survey. BMC Infect Dis. 2019;19:1054. https://doi.org/10.1186/s12879-019-4673-0.

20. Rowley J, Vander Hoorn S, Korenromp E, Low N, Unemo M, Abu-Raddad LJ, et al. Chlamydia, gonorrhoea, trichomoniasis and syphilis: global prevalence and incidence estimates, 2016. Bull World Health Organ. 2019;97(8):548.