CHAPTER 8

PERFORMANCE STUDY OF PROPOSED MODEL AND EXISTING MODELS ON HEART DATASETS

In this chapter, the proposed model, comprising of ModifiedBoostARoota (MBAR) feature selection algorithm and Optimized Super Learner Ensemble Model (OSLEM) classifier consisting of CatBoost (CatB) and Decision Tree (DT) as base models, is compared with various existing models experimented on heart datasets. For the comparative study, two low dimensional and high dimensional datasets are considered. The model evaluation metrics are used to compare and contrast the models.

In the existing related works, a variety of ensemble models and hybrid models are devised with different combinations of the individual classifiers namely, RF, LR, SGD, SVM, Extreme Learning Machine (ELM), KNN, Gradient Boosting classifier (GBC), AdaB, ANN, XGB and NB. The next sections compare and contrast the suggested model's performance with those of similar works.

8.1 Evaluation of Proposed Model and Related Works on Cleveland Heart Dataset

Table 8.1 contrasts the proposed model (MBAR with OSLEM) with the existing models assessed on the Cleveland heart dataset in terms of model evaluation metrics. In Table 8.1, on comparing the precision values, the proposed model (MBAR with OSLEM) gives a precision score of 95.8%, which is very high compared to other models' values. The recall value of the existing work BHO Algorithm with XGBoost (Rajadevi, 2021) is 92.0% which is much higher than all other works in comparison. The proposed model has an F1-score that is higher than most others (92%). The suggested model has an accuracy of 93.40%, placing behind it is the 92% accuracy of models BHO with XGB (Rajadevi, 2021) and ReliefF with an ensemble model (RF, LR, SVM, ELM, KNN) (Zhenya & Zhang, 2021). When compared to other models using the Cleveland heart dataset, the proposed model clearly outperforms them in terms of f1-score and accuracy. With 95.8% precision, it's clear that the model is successfully identifying most data. It can be inferred that the model is performing well in terms of overall classification correctness.

Table 8.1 Performance Metrics of Models on Cleveland Heart Dataset

Related Work	Methodology	Precision	Recall	F1-score	Accuracy
Hera et al., 2022	RF, Multi-Tier Ensemble (MTE) (Stacking (RF, LR, SGD), bagging (GBC, ADA))	85.97%	79.72%	79.14%	86.21%
Doppala, 2022	Ensemble (NB, RF, SVM, XGBoost) with voting	85.00%	90.00%	88.00%	88.24%
Rajadevi, 2021	Black Hole Optimization Algorithm (BHO), XGBoost	90.00%	92.00%	91.00%	92.30%
Zhenya & Zhang, 2021	ReliefF, ensemble model (RF, LR, SVM, ELM, KNN)	88.67%	89.68%	88.84%	91.60%
Wijaya et al., 2018	Particle Swarm Optimization (PSO), NB	87.77%	88.67%	88.22%	86.67%
Wenxin, 2020	Ensemble model(DT, SVM, ANN)	82.80%	90.80%	87.00%	87.00%
Proposed model	**MBAR with OSLEM**	95.80%	88.50%	92.00%	**93.40%**

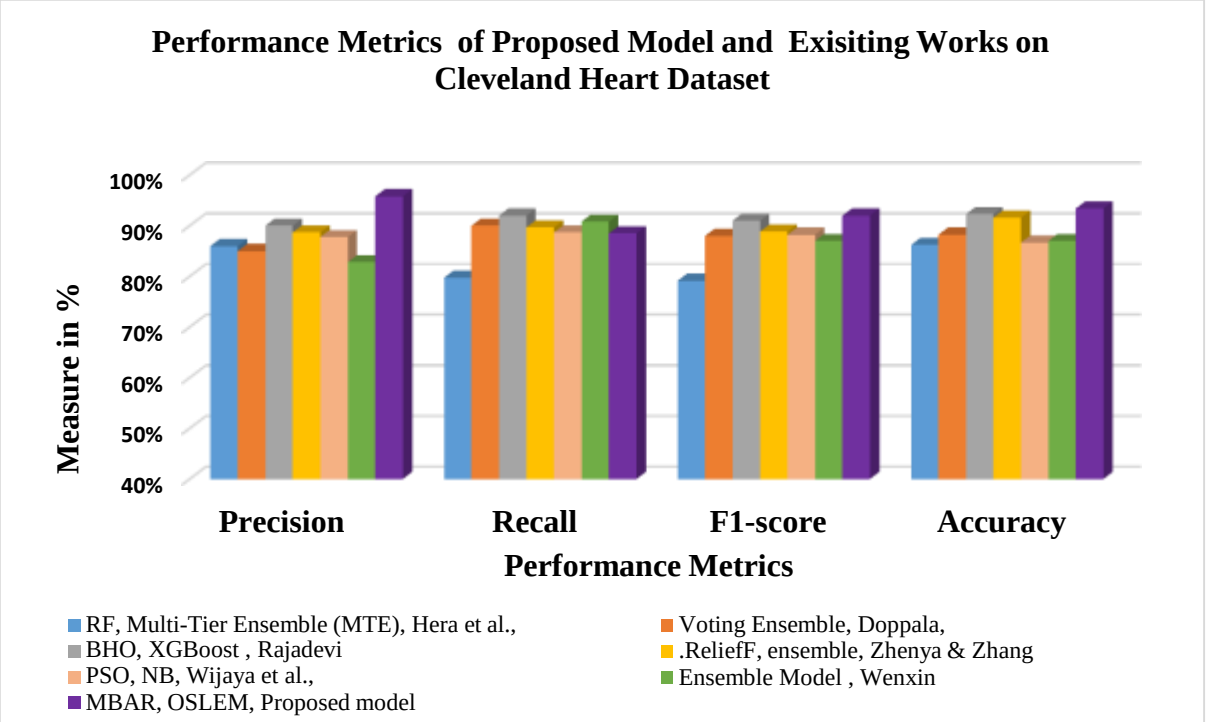

Figure 8.1 Comparison of Proposed Model with Existing Related works on Cleveland Heart Dataset

Figure 8.1 is a visual representation of a comparison between the proposed model and relevant prior art on the Cleveland Heart dataset. Figure 8.1 demonstrates that the suggested model achieves higher levels of accuracy, precision, and f1-score than prior efforts. The high f1-score demonstrates that the suggested model strikes an appropriate balance between precision and recall.

8.2 Evaluation of Proposed Model and Related Works on Statlog Heart Dataset

Table 8.2 presents a comparison of (MBAR with OSLEM) model with the earlier related works on the Statlog heart dataset in terms of model evaluation metrics. According to Table 8.2, the stacking model by Wang et al. (2020) shows high precision compared to other models but the proposed model outperforms others in terms of accuracy, scoring a high at 92.6%. The suggested model achieves a similar F1-score, precision, and recall on the test data, suggesting a comparable amount of false positives and false negatives. Accurately detecting positive examples (precision) and capturing all positive instances (recall) both lie at 89%, indicating that the model strikes a good compromise between the two.

Table 8.2 Performance Metrics of Proposed Model and Related Models on Statlog Heart Dataset

Related Work	Methodology	Precision	Recall	F1-score	Accuracy
Hera et al., 2022	RF, Multi-Tier Ensemble (MTE) (Stacking (RF, LR, SGD), bagging GBC, ADA)	84.9%	79.2%	81.5%	84.1%
Wang et al., 2020	Stacking (GNB, GB, RF, ET, ADB, MLP, XGB)	94.7%	85.8%	89%	90.7%
Alam et al., 2019	Ranking Algorithms (infogain, correlation, ReliefF), RF	84.5%	83.5%	83%	83.5%
Proposed Model	**MBAR with OSLEM**	**88.9%**	**88.9%**	**88.9%**	**92.6%**

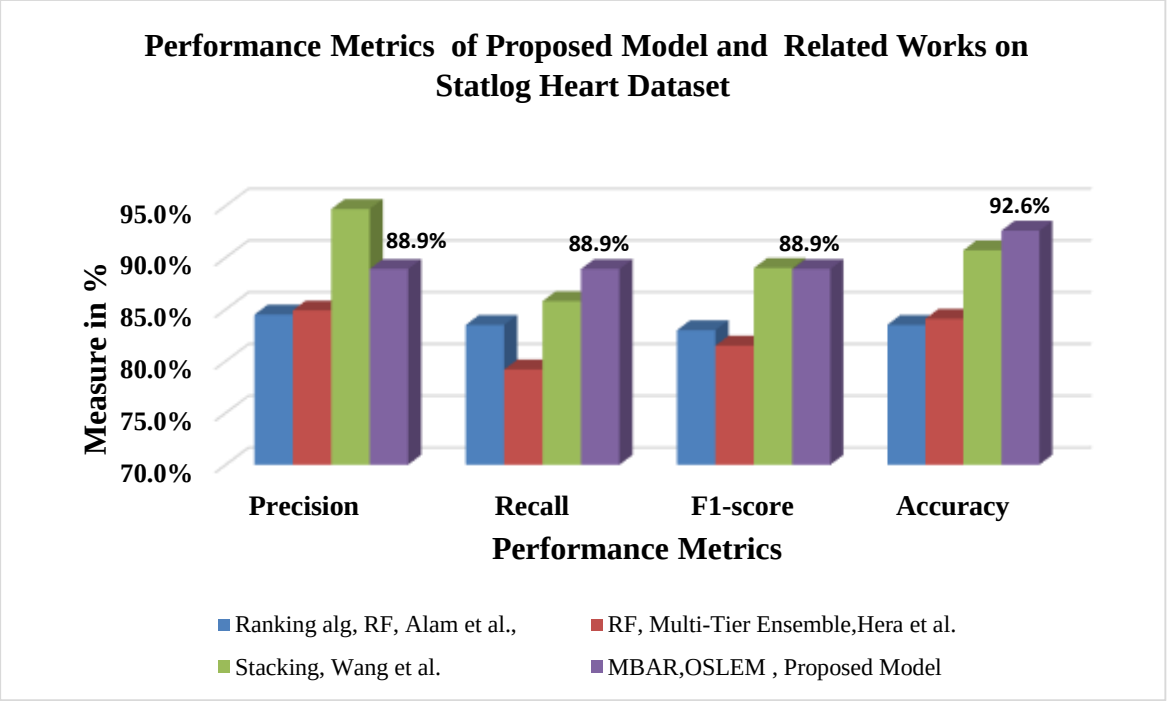

Figure 8.2 Comparison of Proposed Model with Existing Related Works on Statlog Heart Dataset

In Figure 8.2, it can be noticed that the model by Hera et al. (2022) exhibits less recall compared to other models. As in disease prediction, false negatives need to be reduced, it is ideal to have a model with good recall value. Although, the model by Wang et al. (2020) has high precision, the proposed model has high recall and high accuracy compared to the other models considered. The proposed model's reasonable number of correct predictions and low rate of false positives and false negatives are inferred from its balanced precision and recall.

8.3 Evaluation of Proposed Model with Related Works on Arrhythmia Heart Dataset

The Arrhythmia heart dataset is an imbalanced high-dimensional dataset. So, it is ideal to balance the dataset before training a model on the dataset. Reducing the number of features is another way to boost the performance of the models on this dataset. Experimental results on the Arrhythmia heart disease dataset are summarized in Table 8.3, contrasting the proposed model (MBAR with OSLEM) with other evaluated models in terms of specificity, sensitivity, and accuracy. Synthetic Minority Over Sampling Technique (SMOTE) was used to ensure data equality, and then the proposed model was put into action.

PCA is utilized by the majority of researchers (as shown in Table 8.3) for feature selection, and hybrid classifiers (with SVM being the most popular choice) are employed for classification. The suggested model exceeds earlier researchers' models in sensitivity and accuracy, while all other models investigated in this study have performance metric values differing by ±5.

Table 8.3 Performance Metrics of Proposed Model and Related Models onArrhythmia Heart Dataset

Related Work	Methodology	Specificity	Sensitivity	Accuracy
Iyer et al., 2021	Principal Component Analysis (PCA); XGBoost	92.2%	75.8%	83.20%
Pandey et al., 2020	PCA ; SVM, NB	80.0%	70.0%	89.74%
Mitra & Samanta, 2013	Correlation-based Feature Selection (CFS); Levenberg-Marquardt (LM)	88.4%	86.7%	87.71%
Yilmaz, 2013	Fisher score, Least-squares SVM	82.0%	84.86%	82.09%
Abirami & Raj, 2020	SVM, RF	83.3%	81.3%	85.15%
Proposed Model	**MBAR with OSLEM**	88.0%	91.7%	**90.00%**

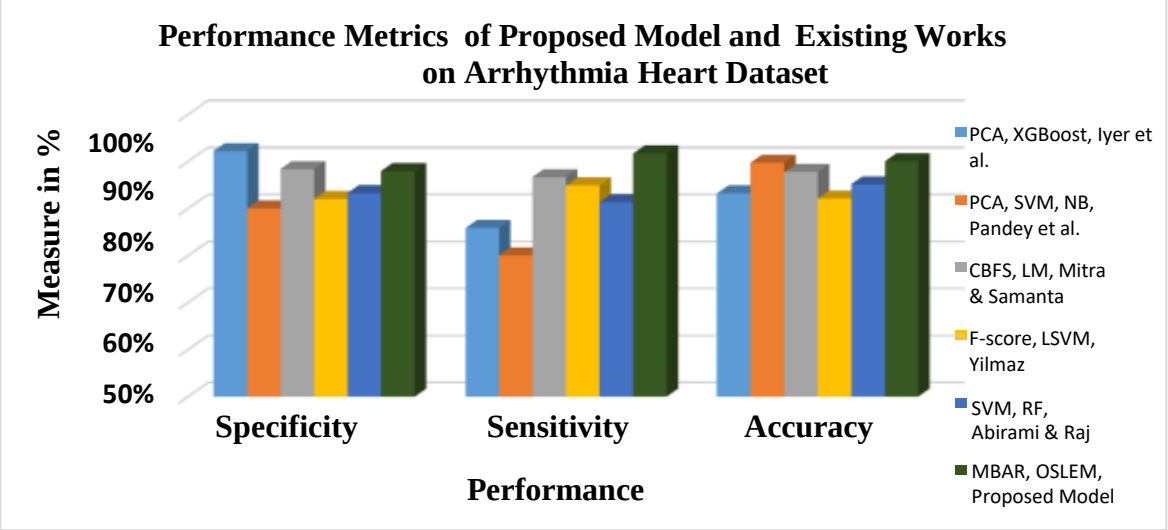

Figure 8.3 Comparison of Proposed Model with Existing Related Works onArrhythmia Heart Dataset

Figure 8.3 is a visual representation of a comparison between the MBAR with OSLEM model and prior work on the Arrhythmia heart dataset. High sensitivity in the suggested model indicates that it has categorized less false negatives than the other models,

as shown by the chart. Although the work by Iyer et al. (2021) has higher specificity, the proposed model is preferred because in disease classification there should be fewer false negatives, that is, high recall is very essential. When compared to competing models, the proposed one also provides superior accuracy.

8.4 Evaluation of Proposed Model with Related Works on Z-Alizadeh Sani Heart Dataset

The performance metrics Specificity, Sensitivity, and Accuracy of the proposed and prior works on the Alizadeh Sani heart dataset are tabulated in Table 8.4. The comparison conveys that the stacking ensemble model (Wang et al., 2020) has performance very close to the proposed model but the proposed model outperforms it by a narrow margin in all the metrics considered. The Genetic Support Vector Machine Along with ANOVA (GSVMA) model (Hassannataj et al., 2022) has the highest specificity of 100% but has low sensitivity among the models considered in this study. It is evident that all the models' performances are on par as they have performance metrics values almost differing only by ±5%.

Table 8.4 Performance Metrics of Proposed Model and Related Models on Z-Alizadeh Sani Heart Dataset

Related Work	Classifier	Specificity	Sensitivity	Accuracy
Arabasadi et al.,2017	Genetic Algorithm, Neural Networks	92.0%	97.0%	93.85%
Kilic et al., 2018	ABC algorithm and SMO	89.43%	89.35%	89.44%
Wang et al.,2020	Stacking(GNB, GB, RF, ET, ADB, MLP, XGB)	94.44%	95.84%	95.43%
Hassannataj et al., 2022	Genetic Support Vector Machine along with ANOVA (GSVMA)	100%	81.22%	89.45%
Proposed Model	**MBAR+ OSLEM**	**97.5%**	**97.87%**	**97.70%**

The model by Arabasadi et al. (2017), has good sensitivity but the proposed model has the highest sensitivity of 97.87%, which is highly required metric in the case of models used for disease prediction.

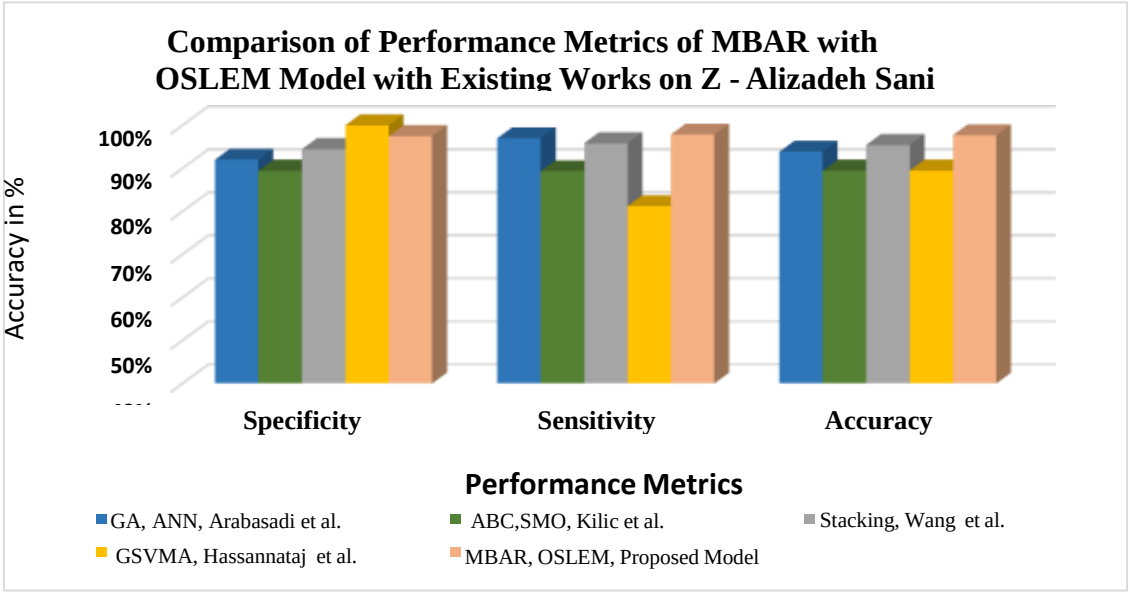

Figure 8.4 Comparison of MBAR with OSLEM with Existing Related Works on Z-Alizadeh Sani Heart Dataset

Figure 8.4 demonstrates that, with the exception of specificity, performance indicators for the proposed model are generally high. Classification results are neutral since the proposed model was trained using a balanced version of the Z-Alizadeh Sani dataset. Another crucial statistic for disease prediction models is sensitivity, and it can be shown that the suggested model has the highest sensitivity score.

8.5 LIMITATIONS

The proposed Machine Learning (ML) model, comprising of ModifiedBoostARoota (MBAR) feature selection algorithm and Optimized Super Learner Ensemble Model (OSLEM) will comprise of the following merits of usual ML models along with the limitations.

- Machine Learning (ML) models works well with structured, smaller datasets (Badawy et al., 2023), in fact, they outperform Deep Learning (DL) models on such datasets (Dong et al., 2022). But for large, complex datasets with unstructured inputs, ML may not be suitable. On such datasets, DL models are used.
- Feature selection is done manually for ML models. So, ML has limitations for tasks requiring feature learning from raw data like Electro Cardio Grams (ECGs), Electronic

Health Records (EHRs) notes, or imaging. For such unstructured data, integration with DL is required because DL can automatically learn relevant features from raw data.

- Stringent data privacy regulations coupled with ethical concerns, create substantial obstacles to collecting and sharing large-scale, multi-institutional datasets.

  To overcome this, Federated Learning, a privacy-preserving machine learning technique can be implemented where data stays on local devices (e.g., hospitals or patient systems) and only model updates are sent to a central server. The central server aggregates these updates to improve the global model, without ever accessing the raw data.

8.6 CHALLENGES

The challenges that are encountered while implementing the proposed ML model in real time to predict heart diseases risk are as follows:

- Structured medical datasets, often derived from Electronic Health Records (EHRs), frequently suffer from missing values due to various reasons like data entry errors and patient non-compliance. This requires complex imputation techniques, which can introduce their own errors and bias in the models.
- Varying data collection protocols across different clinics or hospitals can introduce noise and inconsistencies, significantly impacting model accuracy.
- Seamlessly integrating ML predictions into existing clinical workflows and decision-making processes is a significant practical challenge.
- As healthcare is a critical field, clinicians need to understand the reasoning behind a diagnosis or risk assessment to build trust and take informed decisions. Clinicians may resist ML-based recommendations, especially if they're non-explainable.

8.7 CHAPTER SUMMARY

This section compares the suggested method's performance on the low and high dimension heart datasets to that of similar approaches. The suggested model has a high f1-score and is accurate while still being able to reliably identify true positives. It also shows good performance in terms of overall classification correctness, maintaining a high level of accuracy across all heart datasets. The cost of predicting falsely a heart patient as normal is far higher than that of wrongly labeling healthy individual as heart patient, hence a cost analysis of misclassification helps bring the proposed solution closer to reality. The challenges mentioned can be taken as objectives and addressed in future work.