

Advanced Hybrid and Preprocessing Models for Diagnosis Challenges in Data Classification

Mustafa Adil Fayez * and Sefer Kurnaz

Computer Engineering Department, Science Institute, Altinbas University, Istanbul, Türkiye
Email: mustafaadil302@gmail.com (M.A.F.); sefer.kurnaz@altinbas.edu.tr (S.K.)

*Corresponding author

Abstract—Machine Learning (ML), often viewed as a cutting-edge technology best suited for qualified specialists, presents limited access for other physicians and scientists in the medical profession. In this work, we provide a new, sophisticated, and highly successful technology for medical applications, especially cardiac diagnostics. We propose a novel advanced hybrid optimization model with two essential parts. Initially, we apply a high-performance hybrid resampling technique for feature engineering and pre-processing. This approach, which combines Synthetic Minority Oversampling Technique Edited Nearest Neighbors (SMOTEENN) with Neighborhood Cleaning Rules (NCL), addresses class imbalance in the data. We developed a complex hybrid optimization model that incorporates hyper-parameter optimization, advanced Application Programming Interface (API) functions, and a super-learner ensemble model to enhance diagnosis accuracy in cases where datasets lack balance. Furthermore, we developed high-performance prediction models using sophisticated Support Vector Machines (SVMs). We show that, with re-sampled Cardiovascular Disease (CVD) data, the advanced hybrid optimization model attained an astounding accuracy of 98%. By comparison, an advanced SVM model obtained 96% accuracy, while an advanced deep learning model produced 95.5% accuracy. Our new sophisticated hybrid optimization machine learning models may significantly improve physicians' interpretation of ML results. This strategy could make it easier to apply AI methods on a large scale in the clinic, which would eventually raise patient outcomes and diagnostic accuracy.

Keywords—Cardiovascular Disease (CVD), Neighborhood Cleaning Rules (NCL), Synthetic Minority Oversampling Technique Edited Nearest Neighbors (SMOTEENN), hybrid advanced models, optimization, Application Programming Interface (API) function

I. INTRODUCTION

Machine Learning (ML) is a type of artificial intelligence that uses statistical models to learn from data. Researchers have been paying a lot of attention to it recently because of its ability to optimize models and make predictions that were previously impossible. For instance, this technology enables the prediction of an individual's life expectancy based on their age and other

factors, a feat previously unattainable. Researchers are also using machine learning in cardiovascular disease research to create more accurate models of heart health than ever before. In 2015, at least 102.7 million people in the US population, or 41.5%, had a cardiovascular disease [1]. They estimate that 131.2 million Americans, or 45% of the country's population, will have cardiovascular disease by 2035 [2]. They anticipated that 100 million Americans would suffer from cardiovascular disease by the year 2030. Rather, the country arrived at that benchmark in 2015 and is presently on track to treat in excess of 131 patients with coronary illness by 2035 [3]. Machine learning has seen a huge advancement in the past five years. Clinical analysts have expressed more critical enthusiasm for artificial intelligence calculations. For example, an ongoing study demonstrates that a portion of the medicinal services associations discussed use or want to use advanced machine learning models in imaging [4]. It has become challenging to rely on traditional machine learning systems in the medical fields due to their inability to provide the necessary outcomes and precision needed to effectively deal with the data generated by medical organizations. The objective is to automate the development process of an advanced optimization model that offers competitive performance on a given dataset [5]. This process automates preprocessing, extraction operations, hyper-parameter modification, and method choice. The study will inform diagnostic researchers and professionals about important and developing Machine Learning (ML) techniques. Our discoveries will reveal insight into what advanced machine learning systems can bring, why it is so natural to utilize the apparatus, and how well it performs compared with human expertise [6]. We anticipate that when we test our system on cardiovascular data, the findings will apply to other types of biomedical data. Numerous studies have been conducted in the past to develop machine learning models that can identify similarities in data and consider early identification of cardiovascular disease [7].

This study aims to evaluate the effectiveness and time savings that advanced systems and optimizations offer in comparison to expertly created approaches, with the goal of using advanced ML systems as a solution to the aforementioned issues. Second, we present published studies on cardiovascular disease diagnosis in Part II. We emphasize the significance of advanced Cardiovascular

Disease (CVD) machine learning models and development procedures. Section III then mentions current systems and methods. Section III also presents the patterns observed in high-level implementation. Section V presents the results and solutions for the advanced hybrid optimization system. Section VI is our contribution to the research, and finally, the comparison and conclusion of our study are in Section VII. Sections VIII include a conclusion and a list of references.

II. RELATED WORK

Padmanabhan *et al.* [8] are developing a case study for predicting heart disease using ML that doctors can use in 2019. To prove their claim, they compared auto-ML techniques with those of graduate students using important benchmarks and concluded with the total time needed to train models (ML) and classification accuracy on invisible test datasets. With 70,000 cardiovascular data points, they were 74% accurate.

Mehanović *et al.* [9] uses the Majority Voting Set Technique for Heart Disease Prediction to develop a heart disease prediction model using ANN, KNN, and SVM techniques. Researchers conduct numerous experiments using each of those algorithms. Next, we apply the ensemble's learning and compare the outcomes. In each of these cases, binary classification achieves the highest precision at 87%.

The diagnosis of Coronary Heart Disease (CHD) will be available in 2020, and it will rank important attributes using a random tree model [10]. This system's goal is to improve the diagnostic accuracy of coronary artery disease by selecting important probabilistic features based on grade. They implemented an ensemble approach using machine learning. The findings demonstrate that, in comparison to other models, the random tree model is the most effective identification technique.

Hasan, Najmul, and Bao [11] primarily aims to evaluate the accuracy of various feature selection algorithms in predicting cardiovascular diseases. Experimental results show that using the envelope approximation with the XGBoost classifier can achieve accurate predictions.

Mani *et al.* [12] uses Python programming and machine learning methods, which investigates the prediction of cardiac. To improve the accuracy of CVD diagnosis, the main objective of this project is to develop algorithms to identify whether a person has the condition. They were 86% accurate with their top classifier.

Alfaidi *et al.* [13] suggests a Machine Learning (ML) model for predicting a person's risk of cardiovascular disease. Using publicly accessible cardiovascular disease datasets, seven algorithms were employed to conduct tests and develop models. They were 87% accurate in predicting the disease.

Arroyo *et al.* [14] employed artificial neural networks in this study's training and testing on a CVD dataset with 70,000 cases and 12 variables (ANN). The goal of this study is to enhance ANN performance and improve prediction accuracy.

Alqahtani *et al.* [15] offer a Machine Learning (ML) and Deep Learning (DL) ensemble-based strategy to

predict the chance of a person having cardiovascular disease. The experimental findings showed their model had a disease prediction accuracy of 88.70%.

Bhatt *et al.* [16] develops a model that accurately forecasts cardiovascular disorders to reduce their death rate. We used GridSearchCV to fine-tune the model's parameters to improve the outcome. We tested the suggested model with a dataset of 70,000 cases from Kaggle. This foundational research supports the claim that cross-validated multilayer perceptrons beat all other algorithms in terms of accuracy. We recorded the highest accuracy rate at 87.28%.

Altae and Rad [17] proposed a new hybrid approach for the diagnosis of cardiac disease. This work presented a novel, efficient hybrid model by aggregating the most successful learning classifier with the Particle Swarm Optimization Algorithm (PSOA). Ten trials with a 10-fold Cross-Valuation (CV) allow the suggested hybrid diagnostic approach to record ACC. The findings demonstrate that the suggested approach exhibits an ACC of 93%.

In our research, we are focusing on developing a new advanced hybrid optimization (ML) system that will support a new idea that will be different from the previous one and have high results and performance to handle unbalanced medical data. Furthermore, the majority of previous studies developed their models using 70,000 CVD data and applied renowned machine learning algorithms. The previous models created by previous researchers might not be suitable for big datasets like 70,000 CVD imbalanced data, which contrasts with our study, which developed novel optimization high-performance models for imbalanced data diagnosis problems. In Table I, we have presented the comparison results of this study with those of previous studies.

TABLE I. THE FINDINGS OF THIS STUDY COMPARE PREVIOUS STUDIES

Author	Data	Methods	Accuracy
Padmanabhan <i>et al.</i> [8]	Cardiovascular Data, (70,000) Patient Records	Graduate student	74%
		Auto ML	74%
Alqahtani <i>et al.</i> [15]	Cardiovascular Data, (70,000) Patient Records	Deep learning (DL)	88.70%
Bhatt <i>et al.</i> [16]	Cardiovascular Data, (70,000) Patient Records	Multilayer perceptron	87.28%
Our study with novel advanced hybrid models	Cardiovascular Data, (70,000) Patient Records	Advanced deep learning	95.5%
		Advanced hybrid model	98%
		Advanced SVM	96%

III. METHODOLOGY

In this section of our research, by using (70,000) records of unbalanced cardiovascular disease data, we address sophisticated hybrid optimization machine learning models, such as advanced deep learning, super learners, and advanced SVMs, to create a high-performance system for diagnosing cardiovascular disorders.

A. CVD Disease Dataset

The (70,000) target patient records in the CVD disease dataset represent the presence or absence of a cardiac problem with 12 features, as indicated in Table I [18]. We distinguish three categories of input characteristics: objective, which contains factual information, test, which includes medical experimental outcomes, and subjective, which includes patient information. The target variable “Cardio” contains 35,021 records for Cardiac Patient 0 and 34,979 records for Cardiac Patient 1, out of 70,000 entries in this dataset.

B. Data Preprocessing and Feature Optimization

The process of feature engineering and selection involves coordinating the selection of corresponding functions from the CVD imbalanced dataset, as well as eliminating features that are not crucial for the prediction of the target variable [19]. This improves performance and significantly reduces preparation time. Using low-priority remote CVD data that can have an impact on model predictions, we reprocessed unneeded attributes [20]. We apply our approach to random forest sets and believe that some of the models are excellent for certain data science applications, particularly some classification tasks. We employ this model because it can accept both category and numerical inputs, requiring less pre-processing than some other approaches. We frequently utilize straightforward, pre-built models early in the modeling process to observe how they perform with our data, such as random forests and some other reinforcement techniques [21]. Furthermore, we will employ remedial removal of functionality, a method of greedy optimization that generates a larger subset of features through continuous model creation and evaluation of each subset’s performance.

C. Hybrid Resampling Method

Because many natural processes rarely produce such observations, CVD datasets are imbalanced, making it difficult to distinguish between classes. Cardiovascular disease, for instance, can produce medical data in a population with a limited diagnostic group. Models created by machines can be incorrect [22]. Before the actual analysis, we merged class distribution with data reduction, as we aimed at creating an advanced machine learning system. These new resampling methods were used together with the CVD data set. They are the Neighborhood Cleaning Rule (NCL), humble random and one-sided selection, and Smote-Edited Nearest Neighbor (SMOTEENN) methods [23]. This hybrid resampling model will focus on cleaning the data rather than compressing it by using closest-neighbor approaches. It will also alter the data set by deleting samples that are not “complete” with their surroundings. Identify the closest neighbors of each sample in the sampled class, and if the selection criteria are not satisfied, discard the sample [24]. As of the resampling method, 32,497 records are for Cardio 0 patients, 32,497 records are for Cardio 1 patients due to the NCL model, and 12,408 records are in Cardio 0 out of 13,416 records in Cardio 1 as a result of our hybrid resampling model. Previously, there were 35,021 records

for Cardio 0 patients and 34,979 records for Cardio 1 patients. Figs. 1 and 2 display the findings of our novel hybrid resampling model.

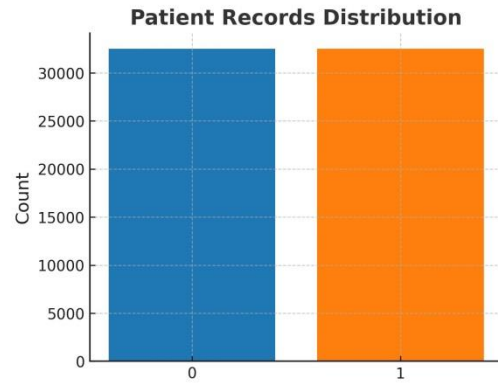


Fig. 1. The results of NCL method.

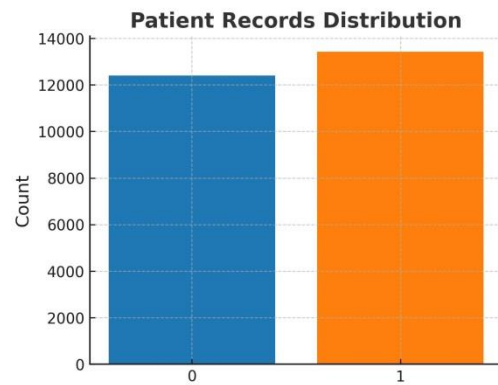


Fig. 2. The results using smote-edited nearest neighbor SMOTEENN method.

Using 10-fold cross-valuation, this procedure evaluates 80 unique hyper-parameter combinations (candidates) generating 800 separate model training and assessment runs. By greatly accelerating the computation using parallel processing with eight concurrent workers, all fits may be finished in sixteen-six-minute times. The updates given at several phases show the effectiveness and advancement of the parallel computing. A model is trained and validated using 10-fold cross-valuation, with 80 candidates and 80 unique hyperparameter combinations evaluated. The model is trained and assessed 10 times, using a new fold as the validation set and the remaining folds as the training set. The backend, (LokyBackend), uses eight CPU cores and eight processes simultaneously. The fitting procedure takes 52.5 s to complete, with a total of 800 fits completed in 16.6 min. The system uses Python parallel computing toolkit Joblib for parallel running. The results of our hybrid resampling approach are shown in Figs. 1 and 2. 32,497 Cardio 0 (CVD-0) patient records and 32,497 Cardio 1 (CVD-1) patient records were used in the sampling, and as a result of our hybrid resampling method using smote-edited nearest neighbors, 12,408 records of (CVD-0) patients and 13,416 records of (CVD-1) patients were obtained (SMOTEENN).

IV. ADVANCED MODELS

A. Advanced Deep Learning Model

In this study, we explore a number of powerful tools that can bring us closer to developing advanced machine-learning models for solving complex problems [25]. We can use the functional Keras API to create advanced deep learning models, share layers from different inputs, and use Keras models [26]. We also covered other more critical activities, such as batch normalization, advanced API functions, and optimization of hyper-parameters.

1) Advanced API function

Tensor manipulation can be performed directly in a functional API. Functions such as layers that accept and return tensors are used [27]. We learned how to use sequences to create patterns. An Application Programming Interface (API) is a more general way to create models than sequential models [28]. It is capable of handling nonlinear topologies, sub-layers, and multi-input or multi-output models.

2) Batch normalization technique

We applied Keras batch normalization to boost our advanced machine-learning models and achieve better results. Before setting the internal parameters of a model, the batch size hyperparameter specifies the size of the sample to process. Batches are one or more sample estimations and for-loop replications [29]. The predictions are compared with the predicted output variables at the end of the sample, and an error is measured [30]. Keras provides batch standardization support via the batch normalization layer. This layer will normalize inputs by transforming them to have 0 mean and 1 standard deviation [31]. This layer will record the statistics of each input variable during training and use those statistics to normalize the data. In the majority of deep learning neural network models, batch normalization is applicable up until a certain point in the model. Before or after the prior layer's activation attributes, the batch normalization layer can normalize the input [32]. The benefits of batch normalization are that the networks train faster, enabling higher learning rates, facilitating the initialization of weights, simplifying the creation of deeper networks, and giving better overall results.

3) Novel hybrid hyper-parameter optimization

Therefore, we must promptly, methodically, and ethically explore the realm of potential decisions. We need to look for architectural spaces and empirically identify the best-performing ones [33]. This is a key area of research that automated hyper-parameter optimization focuses on. The secret to this strategy is that the algorithm uses this history of validation outputs to select the next set of hyperparameters to test, depending on various sets of hyperparameters, learning rates, loss functions, and epochs [34]. It is feasible to use hyper-band receivers, evolutionary algorithms, Bayesian optimization, and simple random search [35]. By combining random forest and Ada-boost techniques with advanced hybrid optimization models, we first used random sampling to increase the performance of our system. We then used Grid Search (GS) with hybrid models and hyper-band

tuners with advanced deep learning to select the best parameters and achieve a good fit. However, we cleverly only continued with the best-performing candidates based on prior results from the system's inception.

B. Advanced Super Learner Model

We will be able to develop and test many Machine Learning (ML) models using a single set of data, allowing the algorithm to combine each model to produce better overall predictions [36]. Superlearning significantly improved behavior recognition models from accelerometers, and researchers have widely used them in geographical analysis and medical research. New (ML) methods may provide a solution to the method selection problem we just created [37]. The process of superlearning entails gathering candidates (other ML approaches), applying them to a dataset, and choosing the best group of learners based on the results of the subsequent cross-validation procedure [38]. In order to produce outputs that are equivalent to any of the individual inputs, the SL looks for the optimal learner combination and uses it. The ML-Ensemble is equally simple to use, and it uses the same general approach to classification issues [39]. Here, we use the ML-ensemble module, together with a subset of classifier models, to create Super Learning as a meta-model for classification.

C. Advanced SVM Model

Recently, the medical community has acknowledged SVM as one of the top techniques for tackling classification issues. A strong mathematical foundation, few tuning parameters, quick classification, and excellent generalization ability are just a few of its many impressive properties [40]. We need to optimize SVM, which includes hyper-parameter tweaking, to create a sophisticated and high-performance machine learning model for CVD diagnosis issues, despite the default parameter not being sufficient [41]. Grid Search is the most effective method we employ for perusing the search space using an equally Spaced Grid (GS) [42]. Kernel, regularization, and gamma are the three basic facets of hyper-parameter tuning. The kernel's primary function is to transform input data from a dataset into the appropriate format. Other types of functions, such as linear, polynomial, and Radial Basis Functions (RBF), RBFs, and polynomials, are helpful for nonlinear hyperplanes. The regularization function maintains regularization in the Python C module, where C is the penalty parameter and denotes the term "error". The SVM optimizer can accept a certain amount of error, depending on the word misclassification or error [43]. As opposed to a higher gamma value, which would result in overfitting, a lower gamma value will roughly match the training dataset. In other words, while high gamma values consider all data points when determining the cutoff, low gamma values only consider nearby points.

V. RESULTS AND DISCUSSION

We use a new hybrid resampling model that combines the Neighborhood Cleaning Rule (NCL) and the nearest

neighbor (SMOTEENN) to get the best results on unbalanced CVD datasets. Additionally, we preprocessed our dataset using methods like feature importance and missing value preprocessing. The SMOTEENN resampling method gave us a distribution of predictable categories (Cardio): 32,497 records for Cardio 0 patients, 32,497 recordings for Cardio 1 patients, 12408 records for CVD-0 patients, and 13,416 records for CVD-1 patients. Performance measurements employed in this study include accuracy, precision, recall, F1-score, and area under the ROC curves (AUC-ROC score). We used a Jupyter notebook running on an HP Z-Book laptop with Python 3 to simulate the proposed model after partitioning the CVD-resampled dataset into two halves.

A. The Results Using Advanced Deep Learning

Hyper-band optimization enabled our advanced deep learning model to boost accuracy by 95%, while batch normalization and API functions enabled us to achieve 95.5%, the highest accuracy possible from an advanced machine learning system. We show the accuracy of the optimization process in Fig. 3.

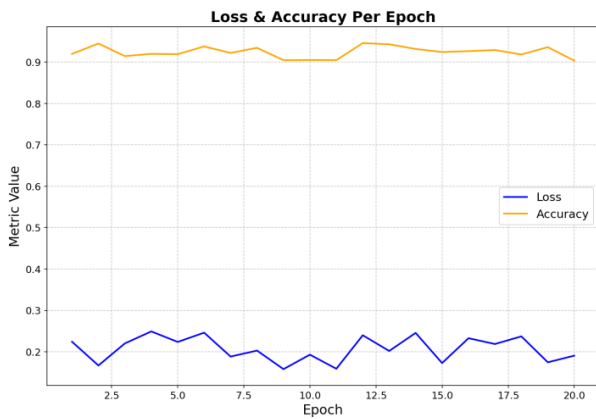


Fig. 3. Optimization accuracy with Re-sampled CVD data and hyper-band model.

The performance and accuracy of our high-level system went up after we used resampled CVD data and high-level API functions to do batch normalization. Fig. 3 above shows the best accuracy of the optimization step, which is about 95%. Fig. 4 depicts the batch normalization outcome.

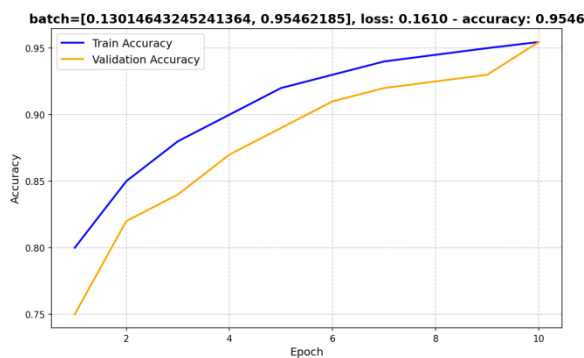


Fig. 4. Batch normalization results curve.

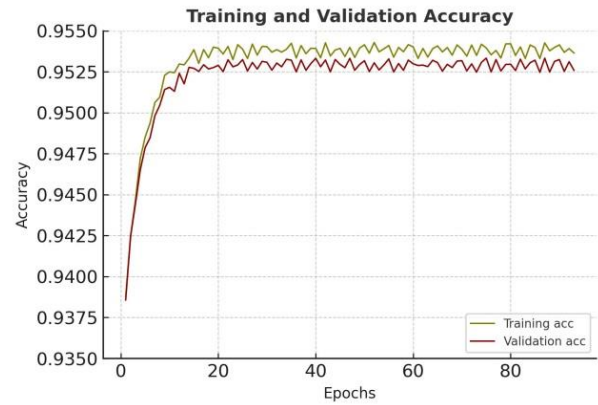


Fig. 5. Accuracy of training and validation in an advanced deep learning model using re-sampled CVD data.

Fig. 5 shows the accuracy of new, enhanced deep learning with batch normalization, high-level API methods, learning rate, batch size, and number of layers. Moreover, we present the training loss and validation loss in Fig. 6.

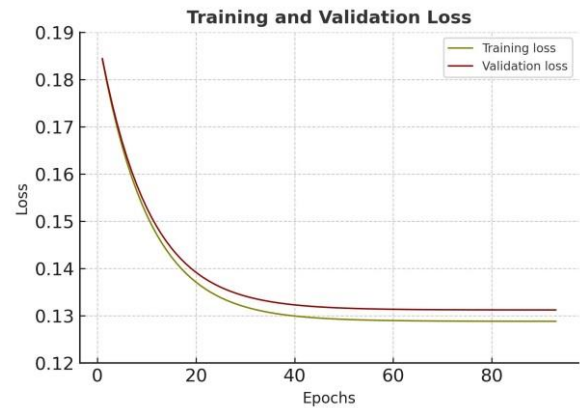


Fig. 6. The loss of training and validation of an advanced deep learning model using re-sampled CVD data.

B. Results of a Super-Advanced Learner

The findings using CVD data records had a 79% accuracy rate, with the train as 34,531 and the test as 14,800. Table II displays the classification report for this model based on the CVD data. Fig. 7 displays the region of the super-learner ROC curve that had a 79% accuracy rate when using re-sampled CVD data sets. Metrics for classifier performance for each classifier under Layer 1, in Table III we display several performance measurements: Mean classifier accuracy score: score-m. The accuracy score's standard deviation, which reflects variability, is known as the score-S. Ft-m represents the classifier's mean fitting time, or training time, expressed in seconds. ft-s: Fit Time Standard Deviation. Calculated in seconds, pt-m is the classifier's mean prediction time—that is, time to predict on the test set. pt-s: prediction time's standard deviation. Super Learner the final performance scores of the super learner—an ensemble model combining the predictions from all the classifiers in layer 1—is 78.635. The figure is probably a mean accuracy percentage, indicating that the great learner achieves an average

accuracy of around 79%. Synopsis: with very short fitting and prediction times, the GaussianNB and Logistic Regression classifiers attain the maximum accuracy (0.79); hence, they are both efficient and effective. Support Vector Classifier (SVC) is less efficient, but it has excellent accuracy because it uses much higher fitting and prediction times. Using the strengths of individual classifiers to maximize performance, Super Learner aggregates all these models to get an overall accuracy of 79%.

TABLE II. THE OUTPUT OF AN INTELLIGENT SUPER-LEARNER MODEL USING RE-SAMPLED CVD DATA

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.78	0.80	0.79	7,347
Cardio 1	0.80	0.77	0.79	7,453
Avg.	0.79	0.79	0.79	14,800

TABLE III. THE OUTPUT OF AN INTELLIGENT SUPER-LEARNER MODEL USING RE-SAMPLED CVD DATA INCLUDING MODELS PROCESSING DETAILS

Models	Score-m	Score-s	Ft-m	Ft-s	Pt-m	Pt-s
Adaboost	78	0.1	4.76	0.15	0.09	0.04
Bagging	76	0.1	5.11	0.29	0.03	0.02
Extratree	76	0.1	0.59	0.04	0.02	0.01
GaussianNB	79	0.1	0.03	0.03	0.00	0.00
Kneighbors	75	0.1	0.06	0.01	5.11	0.51
Logisticreg.	79	0.1	0.20	0.02	0.00	0.00
Randomforest	76	0.1	1.50	0.04	0.02	0.00
SVC	79	0.1	411.84	122.51	2.71	1.04

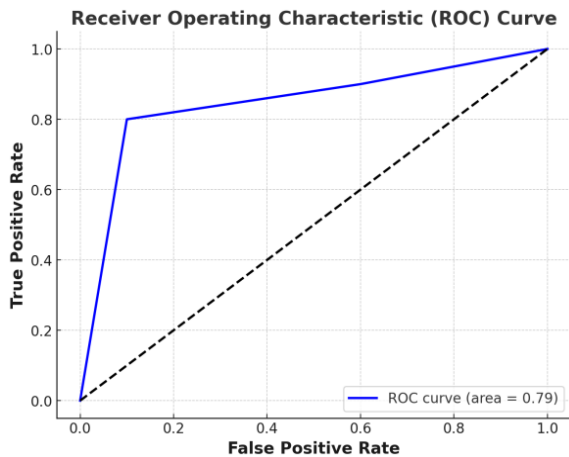


Fig. 7. ROC curve results of super learner model.

C. The Results of Advanced SVM

This part uses SVM hyper-parameter adjustment to create high-performance models. With an evenly spaced grid known as the Search Grid (GS), we employ the most effective methods for searching the search space, including kernel tuning. We used the best set of parameters obtained during development (“C”: 10, “Gamma”: 0.01, “Kernel”: “RBF”) to log CVD data and increased accuracy to 96.43%.

In Table IV, we present the classification results of the tests for the advanced SVM model. The advanced SVM model, when applied to re-sampled CVD data sets, exhibited a 96% accuracy rate, as shown in Fig. 8.

TABLE IV. THE OUTPUT OF AN INTELLIGENT SVM MODEL USING RE-SAMPLED CVD DATA

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.95	0.98	0.96	3,665
Cardio 1	0.98	0.95	0.97	4,039
Avg.	0.96	0.96	0.96	7,704

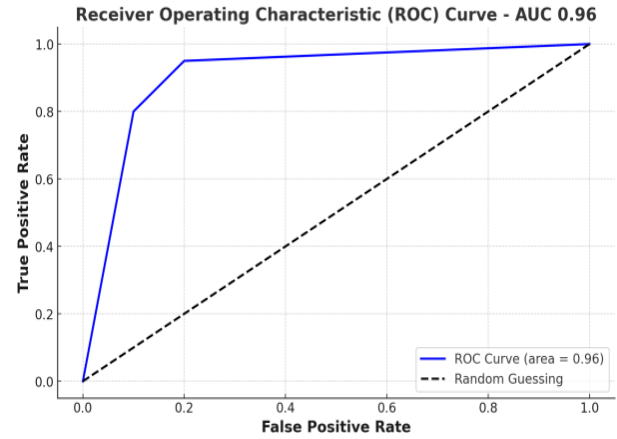


Fig. 8. ROC Curve Display of an Advanced SVM Model with Re-sampled CVD.

D. Advanced Hybrid Optimization Model Results

To obtain high-performance models, we use a hybrid optimization model that includes random forest and Ada boosting in this section. The search grid, which employs a regularly Spaced Grid (GS), is a particularly accurate strategy for exploring the search space. We achieved 98% or greater accuracy with the best set of parameters for CVD data records. Table V displays the classifier report for our novel hybrid optimization model. Fig. 9 displays the area of the ROC curve for the hybrid optimization model, which achieved 98% accuracy by implementing re-sampled CVD data records.

TABLE V. NEW ADVANCED HYBRID OPTIMIZATION MODEL WITH RESULTS ON RE-SAMPLED CVD DATA

Class No.	Precision	Recall	F-Score	Support
Cardio 0	0.98	0.98	0.98	3,732
Cardio 1	0.98	0.98	0.98	4,016
Avg.	0.98	0.98	0.98	7,748

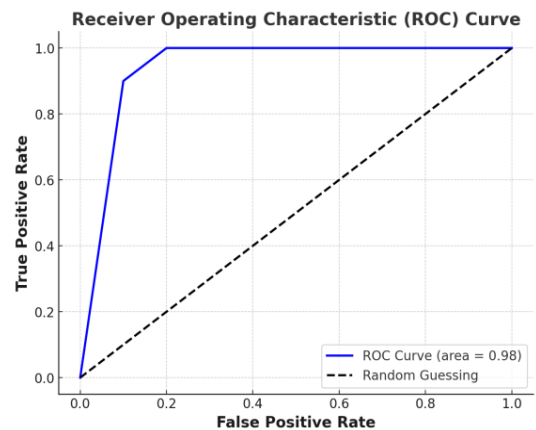


Fig. 9. ROC curves for the advanced hybrid optimization model with re-sampled CVD data.

We developed an advanced and efficient hybrid AI system to handle imbalanced medical data, resulting in highly accurate results. We also present the results in graphs and tables, using specific reports and numbers. We found that better types of machine learning models that deal with imbalanced medical datasets encourage advanced and hybrid approaches, as we have achieved high accuracy rates of 98%. The advanced deep learning optimization model provided the most accurate results, about 95.5% lower than our advanced hybrid optimization model because deep learning techniques require large amounts of data to achieve high accuracy. Also, we see that the accuracy of our results obtained with advanced SVM is 97%, which means that the advanced form of SVM is also a good choice for handling imbalanced datasets. In the end, we obtained 79% accuracy from the Super Learner model, which is lower than other models due to the underlying technology. The Super Learner model generates a powerful learning model that is based on several existing models. Our models yield close results because we are developing them in a futuristic, novel, and advanced form. This includes the novel advanced hybrid re-sampling and preprocessing models, which provide us with the most accurate results of any other study in a shorter amount of time. In this regard, the confusion matrix outcomes in Fig. 10 and the resulting metrics reveal that the model is doing really well in terms of target condition (CVD) diagnosis. It shows a dependable and efficient model fit for medical diagnostics, where both precision and accuracy are critical. In Fig. 11, we display the precision of different classes for an advanced hybrid optimization model.

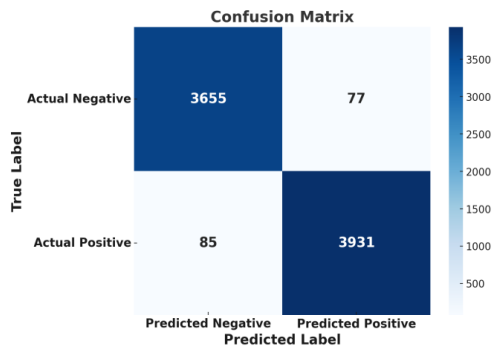


Fig. 10. Confusion Matrix of an Advanced Hybrid Optimization Model with Re-sampled CVD Data.

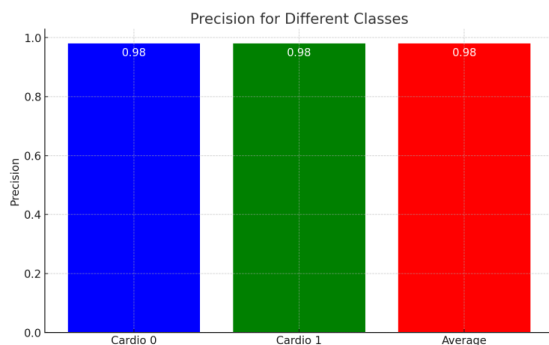


Fig. 11. Precision of an Advanced Hybrid Optimization Model with Re-sampled CVD Data.

VI. RESEARCH CONTRIBUTION

Our research adds the following contributions:

We apply high-performance preprocessing techniques for CVD datasets, including a hybrid resampling method that combines Neighbor Cleaning Rules (NCL) and smote-edited nearest neighbors (SMOTEENN).

Intelligent cardiovascular diagnostic systems also use a resampling approach to test new advanced deep learning techniques like hyper-parameter optimization methods, advanced super learners, and advanced SVM models. Additionally, we are developing a new advanced hybrid optimization model to improve the performance of our medical diagnostic system.

Moreover, we contrast the effectiveness of our novel technological model with another system that makes use of Artificial Intelligence (AI) and computer programming techniques. Finally, we have detailed experimental results on the cardiovascular dataset.

VII. COMPARISON DESCRIPTION

We use three original study papers [8, 15, 16] to compare general trends. The experimental design of the baseline research and the proposed experiments are nearly identical. This study used a novel hybrid optimization method to detect cardiovascular disease. Padmanabhan *et al.* [8] conducted a case study on the risk assessment of cardiovascular disease. They investigated a scenario in which they predicted cardiovascular disease risk. They obtained 85% of the 70,000 cardiovascular disease data by using the Heart UCI data gathering method. Alqahtani *et al.* [15]. They proposed an ensemble-based approach combining Machine Learning (ML) and Deep Learning (DL). According to the experimental findings, their model was accurate in predicting diseases by 88.70%, together with Patel *et al.* [16]. We tested the suggested model using a dataset of 70,000 cases from Kaggle. This foundational study's findings, which demonstrate that cross-validated Multilayer Perceptron (MLP) beats all other methods in terms of accuracy, lead to this conclusion. We achieved the best accuracy rate of 87.28%. Our research focuses on creating a novel, advanced hybrid optimization machine learning system to support a novel concept that is superior to an earlier concept in terms of performance and accuracy. High performance can handle unbalanced medical data. We achieved a better accuracy of 95.5% using the new advanced deep learning model, 96% using the advanced SVM, and 98% using the new hybrid optimization model. In Table V, we present the results of this study in comparison with other recent studies.

VIII. CONCLUSION AND FUTURE PLAN

The primary aim of this work was to introduce new sophisticated hybrid optimization ML models for scientific application, therefore challenging the idea that Machine Learning (ML) is exclusively available to highly experienced individuals. We created a hybrid resampling method that combined SMOTE Edited Nearest Neighbors (SMOTEENN) with the Neighborhood Cleaning Rule

(NCL) to produce a balanced Cardiovascular Disease (CVD) dataset. Tested on this re-sampled CVD data, our sophisticated hybrid optimization model showed better results and accuracy than previous research. The model has a super learner model, advanced SVMs, hyper-parameter tuning, the best learning rates, the best layer numbers, batch normalization, and an advanced deep learning part. These enhancements solved the problems with unbalanced data in CVD diagnosis and produced notable performance increases. Our results show that the more sophisticated hybrid machine learning models we provide can help medical practitioners quickly and economically gain more precise diagnostic information for CVD. Further study of these hybrid models' resilience, generalizability, and interpretability will be the main goal of future work. To improve the diagnostic model's efficiency even more, we want to use a Genetic Algorithm (GA) to optimize parameters.

CONFLICT OF INTEREST

The authors declare no conflict of interest.

AUTHOR CONTRIBUTIONS

The authors' contributions were as follows: Mustafa Adil Fayeze provided the initial draft, conceptualization, data curation, technique, software, validation, and visualization, while Sefer Kurnaz provided conceptualization, supervision, research management, writing, and editing. All authors had approved the final version.

REFERENCES

- [1] J. A. Leopold, B. A. Maron, and J. Loscalzo, "The application of big data to cardiovascular disease: Paths to precision medicine," *J. Clin. Invest.*, vol. 130, no. 1, pp. 29–38, Jan. 2020.
- [2] S. S. Virani *et al.*, "Heart disease and stroke statistics—2020 update," *Circulation*, Jan. 2020.
- [3] Cardiovascular disease may incur \$1 trillion in costs by 2035. (February 2020). [Online]. Available: <https://healthitanalytics.com/news/cardiovascular-disease-may-incur-1-trillion-in-costs-by-2035>
- [4] H. Hacilar, Y. Gormez, B.-G. Burcu, C. Gezer, Z. Aydin, and V. C. Gungor, "Classification of cardiovascular artery diseases using artificial neural network," *Int. J. Biosci. Biochem. Bioinforma.*, vol. 10, no. 1, pp. 34–41, 2020.
- [5] N.-S. Tomov and S. Tomov, "On deep neural networks for detecting heart disease," arXiv preprint, arXiv:1808.07168, 2018.
- [6] S. Parida *et al.*, "Innovation in medicine and healthcare systems, and multimedia," in *Proc. KES-InMed-19 and KES-IIMSS-19 Conferences*, 2019.
- [7] F. I. Alarsan and M. Younes, "Analysis and classification of heart diseases using heartbeat features and machine learning algorithms," *J. Big Data*, vol. 6, no. 1, 2019.
- [8] M. Padmanabhan, P. Yuan, G. Chada, and H. V. Nguyen, "Physician-friendly machine learning: A case study with cardiovascular disease risk prediction," *J. Clin. Med.*, vol. 8, no. 7, 1050, 2019.
- [9] D. Mehanović, Z. Mašetić, and D. Kečo, "Prediction of heart diseases using majority voting ensemble method," in *Proc. the International Conference on Medical and Biological Engineering*, 2019, vol. 73, no. 5, pp. 159–163.
- [10] J. H. Joloudari *et al.*, "Coronary artery disease diagnosis ranking the significant features using a random trees model," *Int. J. Environ. Res. Public Health*, vol. 17, no. 3, pp. 1–24, 2020.
- [11] N. Hasan and Y. Bao, "Comparing different feature selection algorithms for cardiovascular disease prediction," *Health Technol. (Berl.)*, vol. 11, no. 1, pp. 49–62, 2021.
- [12] B. S. S. Mani and V. M. Manikandan, "Heart disease prediction using machine learning," *Handb. Res. Dis. Predict. Through Data Anal. Mach. Learn.*, pp. 373–381, 2020.
- [13] A. Alfai, R. Aljuhani, B. Alshehri, H. Alwadei, and S. Sabbeh, "Machine learning: Assisted cardiovascular diseases diagnosis," *Int. J. Adv. Comput. Sci. Appl.*, vol. 13, no. 2, pp. 135–141, 2022.
- [14] J. C. T. Arroyo and A. J. P. Delima, "An optimized neural network using genetic algorithm for cardiovascular disease prediction," *J. Adv. Inf. Technol.*, vol. 13, no. 1, pp. 95–99, 2022.
- [15] A. Alqahtani, S. Alsubai, M. Sha, L. Vilcekova, and T. Javed, "Cardiovascular disease detection using ensemble learning," *Comput. Intell. Neurosci.*, vol. 2022, 2022.
- [16] P. L. M. Chintan, M. B. Patel, and T. Ghetia, "Effective heart disease prediction using machine learning techniques," *Algorithms*, vol. 16, no. 2, 2023.
- [17] A. A. Altae and A. E. Rad, "Diagnosing heart disease by a novel hybrid method: Effective learning approach," *Informatics Med. Unlocked*, vol. 40, no. March, 101275, 2023.
- [18] S. Ulianova. (February 2020). Cardiovascular disease dataset. [Online]. Available: <https://www.kaggle.com/sulianova/cardiovascular-disease-dataset>
- [19] J. Gatto, R. Lanka, Y. Iwashita, and A. Stoica, "Single sample feature importance: An interpretable algorithm for low-level feature analysis," arXiv preprint, arXiv:1911.11901, 2019.
- [20] W. L. Cava, C. Bauer, J. H. Moore, and S. A. Pendergrass, "Interpretation of machine learning predictions for patient outcomes in electronic health records," in *Proc. AMIA Annual Symposium*, Mar. 2019, 572.
- [21] Q. Zhou, H. Zhou, and T. Li, "Cost-sensitive feature selection using random forest: Selecting low-cost subsets of informative features," *Knowledge-Based Syst.*, vol. 95, pp. 1–11, 2016.
- [22] H. Fujita and A. Selamat, "Advancing technology industrialization through intelligent software methodologies, tools and techniques," in *Proc. the 18th International Conference on New Trends in Intelligent Software Methodologies, Tools and Techniques (SoMeT_19)*, IOS Press, 2019, vol. 318.
- [23] J. Laurikkala, "Improving identification of difficult small classes by balancing class distribution," *Lect. Notes Comput. Sci. (including Subser. Lect. Notes Artif. Intell. Lect. Notes Bioinformatics)*, vol. 2101, pp. 63–66, 2001.
- [24] M. Janicka, M. Lango, and J. Stefanowski, "Using information on class interrelations to improve classification of multiclass imbalanced data: A new resampling algorithm," *Int. J. Appl. Math. Comput. Sci.*, vol. 29, no. 4, pp. 769–781, 2020.
- [25] M. Sewak, M. R. Karim, and P. Pujari, *Practical Convolutional Neural Networks: Implement Advanced Deep Learning Models Using Python*, Packt Publishing Ltd., 2018.
- [26] M. Saha, C. Chakraborty, I. Arun, R. Ahmed, and S. Chatterjee, "An advanced deep learning approach for Ki-67 stained hotspot detection and proliferation rate scoring for prognostic evaluation of breast cancer," *Sci. Rep.*, vol. 7, no. 1, pp. 1–14, 2017.
- [27] A. Gulli and S. Pal, *Deep Learning with Keras: Implement Neural Networks with TensorFlow and Keras*, Packt Publishing, 2017.
- [28] N. K. Manaswi, "Deep Learning with applications using python," *Deep Learn. with Appl. Using Python*, pp. 31–43, 2018.
- [29] J. Wang, S. Li, Z. An, X. Jiang, W. Qian, and S. Ji, "Batch-normalized deep neural networks for achieving fast intelligent fault diagnosis of machines," *Neurocomputing*, pp. 1–13, 2018.
- [30] S. Santurkar, D. Tsipras, A. Ilyas, and A. M. A. Mit, "How does batch normalization help optimization?" *Advances in Neural Information Processing Systems*, vol. 31, 2018.
- [31] A. Gulli, A. Kapoor, and S. Pal, *Deep Learning with TensorFlow 2 and Keras: Regression, ConvNets, GANs, RNNs, NLP, and more with TensorFlow 2 and the Keras API*, Packt Publishing Ltd., 2019.
- [32] S. Wu, G. Li, L. Deng, L. Liu, and D. Wu, "L1-Norm batch normalization for efficient training of deep neural networks," *IEEE Trans. Neural Networks Learn. Syst.*, pp. 1–9, 2018.
- [33] L. Li, K. Jamieson, G. DeSalvo, A. Rostamizadeh, and A. Talwalkar, "Hyperband: A novel bandit-based approach to hyperparameter optimization," *J. Mach. Learn. Res.*, vol. 18, pp. 1–52, 2018.

- [34] M. Feurer and F. Hutter, "Hyperparameter optimization," *Automated Machine Learning: Methods, Systems, Challenges*, pp. 3–33, 2019.
- [35] M. Zheng, W. Tang, and X. Zhao, "Hyperparameter optimization of neural network- driven spatial models accelerated using cyber-enabled high-performance computing," *Int. J. Geogr. Inf. Sci.*, vol. 00, pp. 1–32, 2018.
- [36] M. A. Ladds, A. P. Thompson, J. P. Kadar, D. Slip, D. Hocking, and R. Harcourt, "Super machine learning: Improving accuracy and reducing variance of behaviour classification from accelerometry," *Anim. Biotelemetry*, vol. 5, no. 1, pp. 1–9, 2017.
- [37] A. P. Keil and J. K. Edwards, "You are smarter than you think: (super) machine learning in context," *Eur. J. Epidemiol.*, vol. 33, no. 5, pp. 437–440, 2018.
- [38] A. I. Naimi and L. B. Balzer, "Stacked generalization: An introduction to super learning," *Eur. J. Epidemiol.*, vol. 33, no. 5, pp. 459–464, 2018.
- [39] R. Pirracchio and M. Carone, "The balance super learner: A robust adaptation of the Super Learner to improve estimation of the average treatment effect in the treated based on propensity score matching," *Stat. Methods Med. Res.*, vol. 27, no. 8, pp. 2504–2518, 2018.
- [40] S. Mirjalili, H. Faris, and I. Aljarah, "Evolutionary machine learning techniques," *Cham, Switzerland: Springer*, 2020.
- [41] L. Ali *et al.*, "An optimized stacked support vector machines based expert system for the effective prediction of heart failure," *IEEE Access*, vol. 7, no. c, pp. 54007–54014, 2019.
- [42] H. A. Fayed and A. F. Atiya, "Speed up grid-search for parameter selection of support vector machines," *Appl. Soft Comput. J.*, vol. 80, pp. 202–210, 2019.
- [43] A. Rojas-Dominguez, L. C. Padierna, J. M. Carpio Valadez, H. J. Puga-Soberanes, and H. J. Fraire, "Optimal hyper-parameter tuning of SVM classifiers with application to medical diagnosis," *IEEE Access*, vol. 6, no. October, pp. 7164–7176, 2017.

Copyright © 2024 by the authors. This is an open access article distributed under the Creative Commons Attribution License ([CC BY-NC-ND 4.0](https://creativecommons.org/licenses/by-nc-nd/4.0/)), which permits use, distribution and reproduction in any medium, provided that the article is properly cited, the use is non-commercial and no modifications or adaptations are made.