International Journal of Scientific Research in Engineering and Management (IJSREM)
Volume: 06 Issue: 04 | April - 2022 Impact Factor: 7.185 ISSN: 2582-3930

PREDICTION OF HEART DISEASE USING HYPER PARAMETER TUNING ON RANDOM FOREST CLASSIFIER

Abhishek Ramesh Pardeshi, Rohit Laxman Patil, Sanket Pundlik Patil, Anish Dnyaneshwar Shinde, Kanchan B. Dhomse

MET's Bhujbal Knowledge City

Savitribai Phule Pune University

Adgaon, Nashik-422003, India

ABSTRACT:

Machine learning algorithms are frequently used in the healthcare business to anticipate fatal illnesses. The purpose of this research was to compare the performance of a conventional system with a suggested system that predicts cardiac illness using a Random Forest classification model. Using the grid search technique to the Random Forest classification algorithm, the proposed system assisted in tuning the hyperparameters. The performance of the heart disease prediction system is the major focus of the research. The hyperparameter tweaking model may be used to boost the performance of prediction models. The classic and new approaches were evaluated in terms of accuracy, precision, recall, and F1 score. The proposed hyperparameter tuning model achieved accuracies in the range of 85.25 percent to 91.80 percent, The old approach has accuracies ranging from 81.97 percent to 90.16 percent. When compared to the old technique, these tests demonstrated that the suggested prediction approach is capable of delivering more accurate findings in predicting heart disease with the acquisition of achievable performance.

Keywords: Hyperparameter Tuning, Machine Learning, Random Forest, Disease Prediction, Heart disease Prediction, dataset.

1. INTRODUCTION:

Machine learning is an emerging issue in the healthcare industry for identifying disease and diagnosis, discovering medications, and classifying medical images. The disease prediction system could be useful for hospital administrators, medical practitioners, doctors, and physicians, as well as nursing and residential care facilities.

The death rate from cardiovascular disease is rising at an alarming rate every year. According to a World Health Organization (WHO) estimate, cardiovascular disease was responsible for 31% of deaths globally in 2016. Another analysis from 2015 estimated that 17.7 million people died as a result of cardiovascular disease, such as heart attacks and strokes. When contemporary medical equipment is unavailable, early detection and prediction of cardiac disease are more difficult. Many researchers are aiming to build a heart disease prediction model for early therapeutic intervention using the Random Forest machine learning algorithm.

People anticipate the greatest level of treatment facilities and services in the healthcare sector, which is one of the most significant industries. The purpose of this study is to improve the performance of machine learning systems by using grid search. The grid search may be used to find the best parameters for machine learning systems . The performance of the heart disease prediction system can be improved by using these fine-tuned hyperparameters. The contributions to the research study are listed below .

- In the first step, the authors used Random Forest (RF) algorithms to develop a typical heart disease prediction model.
- In the second phase, the authors proposed a prediction system using a hyperparameter tuning technique and a machine algorithm (RF). To determine the best hyperparameters for each method, a Grid search is used.
- Finally, in terms of accuracy, precision, recall, and F1 score, the standard state was utilised to compare the performance of these two systems.

2. LITERATURE REVIEW:

This section of the project discusses the project's theoretical basis, beginning with an explanation of heart illness, then moving on to overviews of machine learning-related work, and lastly to Heart Disease diagnosis challenges. There are a number of publications available that provide information on how to diagnose heart disease. Here are some of the papers that were mentioned in the references.

This section mostly summarises the contributions of existing cardiac disease prediction tools. Researchers have created a number of machine learning classification algorithms to predict heart disease datasets.

The research paper [1] offered a system that used a Random Forest & ANN algorithms to detect the heart disease in early stages. This model obtained a 85.25 percentage of accuracy. The research [2] described a method that combines two SVM models for accurate heart disease prediction, with accuracies ranging from 57.85 percent to 91.83 percent. The

paper [3] used decision trees, random forests, SVMs, neural networks, and LRs as machine learning classifiers. With an AUC of 0.75, SVM was found to be the best classification model. The research [4] offered a framework for predicting the heart disease dataset that was implemented using five algorithms: RF, Support Vector Machine, Nave Bayes, Logistic Model Tree & Hoeffding DT. These algorithms correctly detected heart disease with accuracies ranging from 81.24 percent to 95.05 percent after picking the best attributes. The study [5] used a DT to develop a hyperparameter tuning model and tested it on 102 heterogeneous datasets. The prediction models in this study [6] were built using DT, SVM, RF, and LR, with the RF achieving the highest accuracy of 90%. To achieve high predicted accuracy, the study [7] offered a fine-tune prediction model to identify relevant features as well as a classification model that included RF, SVM, and DT models. Ensemble approaches, including as bagging and boosting, were found to be useful in enhancing prediction accuracy in the paper [8]. (85.48 percent). [9] A review paper described numerous machine learning research works in the prediction of cardiac disorders. In the publication [10], the authors suggested a system and compared DT, SVM, RF, and LR with selected and full characteristics. From 1992 to 2019, all relevant studies based on heart disease diagnosis were summarised in the review publication [11]. To predict cardiovascular disease, the paper [12] used k-NN, DT, Naive Bayes, LR, SVM, and Neural Network To pick ideal characteristics and improve accuracy, the study [13] created a system including DT, LR, SVM, Nave Bayes, and RF classifiers. The prediction model for heart disease with the hybrid random forest with a linear model produced an improved performance level with an accuracy level of 88:7% in the research publication [14]. The research [16] used the DT, LR, SVM, MLP, Nave Bayes, and RF classifiers to identify optimal features and enhance accuracy in a heart disease prediction system. Using Cleveland and statlog project heart datasets, the study [17] offered a model to predict the classification model and to know which selected features play a crucial role in the prediction of heart disease. Using an adaptive neuro-fuzzy inference system, the study [18] created a method for classifying patient-based characteristic data for heart disease severity. The study [19] created a hybrid intelligent system architecture for the prediction of heart illness using several machine learning classifiers. The authors of the research [20] proposed a cardiac prediction model that recognised and reduced noise. XGBoost was created as a way to predict heart disease in the study.

The researchers created a heart disease prediction system that does not require hyperparameter adjustment. As a result, we proposed a system for predicting cardiac disease utilising several machine learning algorithms and a hyperparameter tuning

strategy.

3. MATERIALS AND METHODS:

The study materials and methodologies are presented and briefly explored in this part.

3.1 Dataset Description

This work uses the Cleveland Heart Disease dataset [1-3] from the UCI machine learning repository for both training and testing purposes. There are 303 instances and 75 attributes in all, but this study only looks at a selection of 14 numerically valued attributes. The output level is divided into two categories, with 0 indicating no heart disease and 1 indicating cardiac illness. Table 1 contains information about the heart disease dataset. The minimum, maximum, mean, and standard deviation are all presented, as well as the attribute's name, description, and minimum, maximum, mean, and standard deviation.

Table 1. *Heart disease dataset description*

Attribute Name	Attribute Description	mean	min	max
age	age in years	54.37	29.00	77.00
sex	1: male, 2: female	0.68	0.00	3.00
Chest pain	1: typical angina, 2: typical type angina, 3: non angina pain, 4: asymptomatic	0.97	0.0	3.0
thestbps	resting blood pressure (in mm Hg)	131.62	94.0	200.0
chol	serum cholestoral in mg/dl	246.26	126.0	564.0
fbs	fasting blood sugar >120 mg/dl), 1 = true; 0 = false	0.15	0.0	1.0

restecg	0: Nothing to note 1: ST-T Wave abnormality 2: Possible or definite left ventricular hypertrophy	0.53	0.0	2.0
thalach	maximum heart rate achieved	149.65	71.0	202.0
exang	exercise induced angina (1 = yes; 0 = no)	0.33	0.0	1.0
oldpeak	ST depression : continuous value	1.04	0.0	6020
solpe	0: Upsloping: better heart rate with exercise 1: Flatsloping: minimal change 2: Downslopins: signs of unhealthy heart	1.40	0.0	6.20
ca	number of major vessels (0-3) colored by flourosopy	0.73	0.0	4.0
thal	1,3: normal 6: fixed defect: used to be defect but ok now 7: reversible defect: no proper blood movement when exercising	2.31	0.0	3.0
target	Heart disease patient=1, healthy= 0			

3.2 Methodology

Two phases of heart disease prediction were included in the study. The classical heart disease prediction system has been given without a machine learning algorithm hyperparameter tweaking technique

3.2.1 Data collection

Medical data is acquired from several sources, such as the patient's medical history, laboratory reports, inquiries, and observations, and preserved as text, a number value, or an image. After gathering data, many academics can use this electronic media raw data to experiment with various healthcare prediction models. The heart disease dataset was obtained from the UCI machine learning repository for this study.

3.2.2 Data preprocessing

Data preparation is used in this step to locate missing values, process noisy, incomplete, irreverent, and inconsistent values, and reduce redundancy in some attributes. Separation, feature scaling, and normalising are then used to determine the data's standard format. Following data preparation, the dataset is separated into a training set (which contains 80% of the data) and a test set (which contains 20% of the data) (20 percent of data).

3.2.3 Model generation (RF)

In this step, machine learning techniques are used to construct multiple categorization models from the training data. Following that test set, individual samples are categorise based on the developed models. Random Forest classifiers, a machine learning method, are used in both the standard and suggested models. The test set is then classified, and the results are assessed using the model that was developed. After that, the regular system is run without any hyperparameter tuning strategies. The categorization model is created using the system's default parameters.

The suggested system is implemented using a grid search approach, and the hyperparameters are tweaked and optimised. Hyperparameter tuning is one of the most important study topics in machine learning. It is thought that if the hyperparameters are tweaked or improved, the machine learning algorithms will perform better. Grid search is a well-known method for fine-tuning the best parameters for various machine learning algorithms. Grid search is an exhaustive search that can be used to find the best hyperparameter values. It can create a model that creates each parameter combination and stores each model combination. This search can help you save time and resources. The RF classifier model is then constructed using the customised parameters. The test set is applied to the proposed model with the tune hyperparameter once the classification model is generated, and the test set's performance is evaluated.

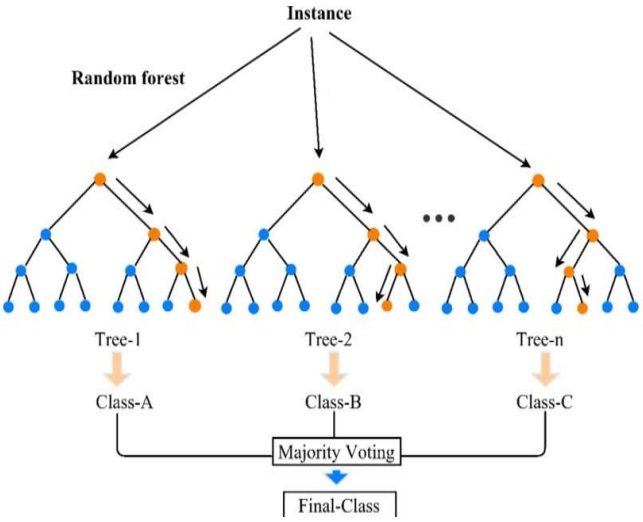
3.2.4 Machine learning algorithm

In the model development process, Random Forest (RF) classifiers are utilised as a machine learning approach. These algorithms were chosen for the heart disease prediction system because they outperformed other machine learning algorithms.

- **Random Forest:**

The Random Forest is a classification technique that uses many decision trees to classify data. When developing each individual tree, it employs bagging and feature randomization to generate an uncorrelated forest of trees whose committee forecast is more accurate than that of any one tree.

Random Forest generates a large number of decision trees and then combines them to provide a more precise and reliable forecast. A decision tree's or a bagging classifier's hyper-parameters are comparable to those of a random forest. As it develops the trees, the random forest adds additional unpredictability to the model.



3.2.5 Hyper Parameter Tuning

Data is used to learn model parameters, and hyper-parameters are tweaked to achieve the best fit. A machine learning model has a lot of parameters to control, and by altering these parameters, the model's performance can be improved. Hyperparameter tuning is the most effective way to evaluate a classifier's effectiveness by executing a large number of parameter combinations. Overfitting is a fundamental machine learning problem that occurs when a classifier is evaluated using training data. When a classifier is assessed using training data, overfitting is a fundamental machine learning problem. When a model performs poorly on test data but very well on rainy data, this is known as overfitting. As a consequence, cross-validation is integrated with the grid search strategy for hyperparameter optimization.

- **Hyperparameter tuning- grid search method:**

Because finding the ideal hyper-parameter can be time-consuming, search algorithms such as grid search and random search are employed. The grid search method is a technique for determining the best classifier parameters so that a model can accurately predict unlabeled input. Some hyperparameters that cannot be learned directly from the training process are tuned using the Grid Search approach. There are numerous hyperparameters in the classification model, and determining the ideal combination of these values is a difficult task. The Grid Search method is one of the most effective approaches for this purpose. Consider the hyperparameters h1, h2, and h3 of a machine learning model X. It will create many different versions of X using all conceivable h1, h2, and h3 combinations. A grid is a collection of hyperparameter values.

3.2.5. Performance Evaluation

In this stage, both systems examine the performance of the training and test sets and calculate the confusion matrix. The confusion matrix was then used to construct and analyse the accuracy, precision, recall, and F1 score metrics for these two models. Accuracy is defined as the proportion of properly recognised observations to the total number of observations. Precision is calculated using the ratio of successfully classified positive samples to total predicted positive samples. Divide the number of positively recognised samples by the total number of samples in the real class yes to get the recall. The F1 score is calculated using a weighted average of accuracy and recall [1][2]. The mathematical concepts are shown in Equations (1), (2), (3), and (4).

Accuracy = (TP+TN) / (TP+TN+FP+FN) --------(1)

Precision = TP / (TP+FP) ----------(2)

Recall = TP / (TP+FN) ----------(3)

F1 score=

2*(Recall*Precision) / (Recall+Precision)

----------(4)

True Positive, False Negative, False Positive, and True Negative, respectively, are denoted by TP, FN, FP, and TN.

4. EXPERIMENTAL RESULTS ANALYSIS AND DISCUSSION:

4.1 Experimental results of the traditional and proposed system

The next sections present the experimental findings of the Random Forest classifier in both the classic and proposed systems.

4.1.1 Performance evaluation and comparison of the traditional system

The machine learning algorithm is used with the default parameters in this experiment. Table 2 shows the output.

In the training phase, the Random Forest (RF) Classifier is fitted with 1000 number of estimators, Gini index, and min samples split=0.2, and the model is run with these parameters, yielding accuracy, precision, recall, and F1 score of 84 percent, 90 percent, 87 percent, and 86 percent, respectively.

Table 2. *Comparison of Performance evaluation on the training set and the testing set.*

Traditional System		Performance evaluation of Training dataset				Performance evaluation of Test dataset			
Machine learning algorithms	**Parameters**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 score (%)**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 score (%)**
RF	criterion= 'gini', min_samples_split=0.2 n_estimators= 1000	100	100	100	100	86	84	90	87

Random Forest is fitted and executed the model with 1000 no. of estimators and Gini index in the final phase of training, yielding accuracy, precision, recall, and F1 score of 100 percent, 100 percent, 100 percent, and 100 percent, respectively. This Random Forest model forecasts the test set with accuracy, precision, recall, and F1 scores of 84 percent, 90 percent, 87 percent, and 86 percent, respectively.

4.1.2 Performance evaluation and comparison of the proposed system

In order to find the optimum hyperparameters, the suggested technique uses Grid search. After the hyperparameters have been fine-tuned, classification models are generated. The proposed system's result is shown in Table 3.

RF is fitted and run in the training phase with tuned hyperparameters of 1000 estimators, entropy index, and 0.3 minimum samples 100 percent accuracy, precision, recall, and F1 score were judged to be 100 percent, 100 percent, 100 percent, and 100 percent, respectively, after the split.

This RF model predicts the test set with 83 percent accuracy, precision, recall, and F1 score, respectively.

Table 3. *Performance of evaluation and comparison of classification model with a hyperparameter tuning approach on the training set and test set*

The proposed system with hyperparameter tuning		**Performance evaluation of Training dataset**				**Performance evaluation of Test dataset**			
Machine learning algorithms	**Tuned hyperparameters**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 score (%)**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 score (%)**
RF	criterion= entropy, min_samples_split=0.3 n_estimators=1000	100	100	100	100	84	83	94	88

4.1.3 Performance evaluation and comparison of the proposed system

In order to find the optimum hyperparameters, the suggested technique uses Grid search. After the hyperparameters have been fine-tuned, classification models are generated. The proposed system's result is shown in Table 3.

RF is fitted and implemented in the training phase with tuned hyperparameters of 2000 estimators, gini index, and 0.4 minimum samples. 100 percent accuracy, precision, recall, and F1 score were found to be 100 percent, 100 percent, 100 percent, and 100 percent, respectively, based on the split. This RF model predicts the test set with 83 percent accuracy, precision, recall, and F1 score of 94 percent, 88 percent, and 84 percent, respectively.

Table 3. *Performance of evaluation and comparison of classification model with a hyperparameter tuning approach on the training set and test set*

The proposed system with hyperparameter tuning		**Performance evaluation of Training dataset**				**Performance evaluation of Test dataset**			
Machine learning algorithms	**Tuned hyperparameters**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 score (%)**	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 score (%)**
RF	criterion= gini, min_samples_split=0.4 n_estimators=2000	100	100	100	100	84	83	94	88

4.1.4 Performance comparison between traditional system and proposed system

Tables 4, 5, 6, and 7 compare the proposed system's performance to that of the existing method in terms of accuracy, precision, recall, and F1 score, respectively.

Table 4. *Comparison of accuracy*

Machine learning algorithms	**Accuracy (%) of Training dataset**		**Accuracy (%) of Test dataset**	
	Without parameter tuning	**With Hyperparameter tuning**	**Without parameter tuning**	**With Hyperparameter tuning**
RF	100	100	80	86

Table 5. Comparison of precision

Machine learning algorithms	Precision (%) of Training dataset		Precision (%) of Test dataset	
	Without parameter tuning	With Hyperparameter tuning	Without parameter tuning	With Hyperparameter tuning
RF	100	100	81	84

Table 6. Comparison of recall

Machine learning algorithms	Recall (%) of Training dataset		Recall (%) of Test dataset	
	Without parameter tuning	With Hyperparameter tuning	Without parameter tuning	With Hyperparameter tuning
RF	100	100	89	90

Table 7. Comparison of F1 score

Machine learning algorithms	F1 score (%) of Training dataset		F1 score (%) of Test dataset	
	Without parameter tuning	With Hyperparameter tuning	Without parameter tuning	With Hyperparameter tuning
RF	100	100	85	87

Tables 4, 5, 6, and 7 compare the performance of the machine learning algorithm Random Forest with and without the hyperparameters tuning technique in terms of accuracy, precision, recall, and F1 score. The findings of these comparisons reveal that prediction systems with hyperparameter customization outperform standard prediction systems.

5. CONCLUSION:

The traditional and new systems were used to forecast the Cleveland heart disease dataset in this paper. Machine learning algorithms, such as Random Forest, are employed in both scenarios to create the heart disease prediction model. These models primarily consist of five phases, although the proposed model differs from the typical system in terms of hyperparameter tweaking. The Random Forest classifiers, on the other hand, provide an accuracy rate of 80 without hyperparameter adjustment. The Random Forest classifiers in the refined set, on the other hand, achieve an accuracy rate of 86 using the hyperparameters tweaking strategy. As a result of the experimental findings of performance evaluation on the heart dataset, it can be inferred that the suggested model is more efficient and can enhance heart disease prediction.

REFERENCES:

[1] A. R. Pardeshi, A. D. Shinde, R. L. Patil, S. P. Patil, K. B. Dhomse," Prediction Of Heart Disease At Early Stages Using Hyperparameter Tuning For Classifier" IRJMET Volume:04/Issue:02/February-2022.

[2] Ali, L., Niamat, A., Khan, J. A., Golilarz, N. A., Xingzhong, X., Noor, A., & Bukhari, S. A. C. (2019). An optimized stacked support vector machines based expert system for the effective prediction of heart failure. *IEEE Access*, Vol.*7*, 54007-54014.

[3] Beunza, J. J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., & Landecho, M. F. (2019). Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *Journal of biomedical informatics*, Vol.*97*, 103257.

[4] Motarwar, P., Duraphe, A., Suganya, G., & Premalatha, M. (2020, February). Cognitive Approach for Heart Disease Prediction using Machine Learning. In *2020 International Conference on Emerging Trends in Information Technology and Engineering (ic-ETITE)*IEEE, 1-5

[5] Mantovani, R. G., Horváth, T., Cerri, R., Vanschoren, J., & de Carvalho, A. C. (2016, October). Hyperparameter tuning of a decision tree

induction algorithm. In *2016 5th Brazilian Conference on Intelligent Systems (BRACIS)* IEEE, 37-42

[6] Ali, A. A. (2019). Stroke Prediction using Distributed Machine Learning Based on Apache Spark. *Stroke*, Vol.*28*, *No.*15, 89-97.

[7] Chandrasegar, T., & Choudhary, A. (2019, March). Heart Disease Diagnosis using a Machine Learning Algorithm. In *2019 Innovations in Power and Advanced Computing Technologies (i-PACT)*, IEEE, Vol. 1, 1-4.

[8] Latha, C. B. C., & Jeeva, S. C. (2019). Improving the accuracy of prediction of heart disease risk based on ensemble classification techniques. *Informatics in Medicine Unlocked*, Vol.*16*, 100203.

[9] Singh, D., & Samagh, J. S. (2020). A COMPREHENSIVE REVIEW OF HEART DISEASE PREDICTION USING MACHINE LEARNING. *Journal of Critical Reviews*, Vol.*7*, *No.*12, 2020.

[10] Ayon, S. I., Islam, M. M., & Hossain, M. R. (2020). Coronary artery heart disease prediction: a comparative study of computational intelligence techniques. *IETE Journal of Research*, 1-20.

[11] Ahmed, H., Younis, E. M., Hendawi, A., & Ali, A. A. (2020). Heart disease identification from patients' social posts, machine learning solution on Spark. *Future Generation Computer Systems*, Vol.*111*, 714-722.

[12] Alizadehsani, R., Abdar, M., Roshanzamir, M., Khosravi, A., Kebria, P. M., Khozeimeh, F., & Acharya, U. R. (2019). Machine learning-based coronary artery disease diagnosis: A comprehensive review. *Computers in Biology and Medicine*, Vol.*111*, 103346.

[13] Amin, M. S., Chiam, Y. K., & Varathan, K. D. (2019). Identification of significant features and data mining techniques in predicting heart disease. *Telematics and Informatics*, *36*, 82-93.

[14] Bashir, S., Khan, Z. S., Khan, F. H., Anjum, A., & Bashir, K. (2019, January). Improving heart disease prediction using feature selection approaches. In *2019 16th International Bhurban Conference on Applied Sciences and Technology (IBCAST)*, IEEE, 619-623

[15] Mohan, S., Thirumalai, C., & Srivastava, G. (2019). Effective heart disease prediction using hybrid machine learning techniques. *IEEE Access*, *7*, 81542-81554.

[16] Wu, C. S. M., Badshah, M., & Bhagwat, V. (2019, July). Heart Disease Prediction Using Data Mining Techniques. In *Proceedings of the 2019 2nd International Conference on Data Science and Information Technology*, 7-11.

[17] Reddy, N. S. C., Nee, S. S., Min, L. Z., & Ying, C. X. (2019). Classification and feature selection approaches by machine learning techniques: Heart

disease prediction. *International Journal of Innovative Computing*, Vol.*9*, *No.*1.

[18] Ziasabounchi, N., & Askerzade, I. (2014). ANFIS based classification model for heart disease prediction. *International Journal of Electrical & Computer Sciences IJECS-IJENS*, Vol.*14*, *No.*02, 7-12.

[19] Haq, A. U., Li, J. P., Memon, M. H., Nazir, S., & Sun, R. (2018). A hybrid intelligent system framework for the prediction of heart disease using machine learning algorithms. *Mobile Information Systems*, *2018*.

[20] Fitriyani, N. L., Syafrudin, M., Alfian, G., & Rhee, J. (2020). HDPM: An Effective Heart Disease Prediction Model for a Clinical Decision Support System. *IEEE Access*, *8*, 133034-133050.